The Influence of Pre-processing on the Estimation of Readability of Web Documents

João Palotti
Vienna University of Techonlogy
Vienna, Austria
palotti@ifs.tuwien.ac.at

Guido Zuccon
Queensland University of Technology
Brisbane, Australia
g.zuccon@qut.edu.au

Allan Hanbury
Vienna University of Techonlogy
Vienna, Austria
hanbury@ifs.tuwien.ac.at

ABSTRACT

This paper investigates the effect that text pre-processing approaches have on the estimation of the readability of web pages. Readability has been highlighted as an important aspect of web search result personalisation in previous work. The most widely used text readability measures rely on surface level characteristics of text, such as the length of words and sentences. We demonstrate that different tools for extracting text from web pages lead to very different estimations of readability. This has an important implication for search engines because search result personalisation strategies that consider users reading ability may fail if incorrect text readability estimations are computed.

Categories and Subject Descriptors: H.3 [Information Systems]: Information Storage and Retrieval
General Terms: Algorithms, Experimentation.
Keywords: Readability, Text pre-processing

1. INTRODUCTION

Search result personalisation is an area of active research within information retrieval [5]. User location, their search history, time the query is issued and type of device used to query are among the many features that current web search engines use to contextualize and personalise the search, aiming to increase the user's satisfaction with the retrieved results. In this paper we investigate one important contextual aspect: the readability of information as presented to users.

The readability of a document is often referred to as the the minimal required level of knowledge to comprehend the text, often measured using the U.S. grade level system. For example, a text with a score of 1 would be suitable for a 6-7 year old child, while a score of 13 requires the knowledge of a freshman undergrad student. In general, the higher the readability score, the harder it is to understand the text.

Within information retrieval, using readability as a way to personalise search results has received substantial attention [3, 15, 18, 9, 17]. For example, Collins-Thompson et al. [3] have investigated methods for estimating user proficiency and readability of results, as well as for re-ranking results according to this information. In health web search, accounting for the readability of the retrieved information is a core requirement to effectively support users (see for example [16]). Health consumers may have a limited understanding of the medical terminology and processes, and thus they should be shown text that is simple to understand and limits expert terminology. Notwithstanding, if experts were to query, the search engine should instead provide more advanced material and detailed information.

Numerous studies have proposed and analysed methods to accurately measure the level of knowledge required to read a text [4]. While recent research has proposed sophisticated readability estimation methods [3, 7], often tailored to specific domains [17], traditional readability measures such as the Automated Readability Index and the Gunning Fog Index are extensively used for assessing information on the web (see for example [16, 18]). These long-established readability measures consider the surface level of the text contained in web pages, that is, the wording and the syntax of sentences. In this framework, the presence of long sentences, words containing many syllables and unpopular words, are all indicators of difficult text to read.

Because traditional readability measures are based on surface level characteristics of text, the accurate parsing of web pages is fundamental to ensure readability is accurately estimated and taken into account for search result personalisation. For example, text contained in different HTML fields, tables, lists, etc., should be adequately processed so as to determine the wording and the syntax of sentences, including sentence length. This pre-processing step is often omitted or simplified (see for example [18]) and the influence of parsing errors on the readability estimation of web pages is unknown. On the other hand, the cleansing of web pages' text has been recognised as an important issue in linguistics and language technology research [1].

In this work we contribute an understanding of:

- how different pre-processing steps influence the estimation of web pages' readability;
- how pre-processing affects the order relations between documents produced by readability measures.

2. READABILITY MEASURES

In this paper we consider a number of traditional readability measures to study how web page pre-processing affects the corresponding estimations of readability. More recent and sophisticated readability measures are not considered

CIKM'15, October 19–23, 2015, Melbourne, Australia.
 ISBN 978-1-4503-3794-6/15/10 ...$15.00.
DOI: http://dx.doi.org/10.1145/2806416.2806613.

here because they usually introduce additional complexity to the estimation process and thus introduce more degrees of variation which are difficult to control for and compare across. For example, Yan's et al. [17] measure requires access to a domain knowledge resource and a method to map text to named entities in such resource.

The readability measures we consider are listed in Table 1; To estimate readability, these measures take into account the surface level of text (i.e., the wording and the syntax of sentences). We do not consider the Dale-Chall and the Flesch Reading Ease (FRE) measures, which share similarities with the measures in Table 1, as their readability score scales widely differ from that of the others, and thus make their comparison more difficult to interpret. An easier way to compare different metrics was also one motivation to update the FRE formula to the FKGL [10].

There are two main factors that have been identified as affecting the user perception of text difficulty and that thus characterise the readability measures of Table 1:

Word Length: short words are commonly used and understood, while long words are usually rare, harder to read, write and remember. This factor is measured by expressions such as $\frac{Sy}{W}$, $\frac{C}{W}$ and $\frac{PW}{W}$.

Sentence Length: while short sentences are usually simple, long ones are usually complex, demanding more cognitive processing and attention. This factor is measured by the average words per sentence, i.e., $\frac{W}{S}$.

Each measure differs from the others in the way these factors are combined, usually via a coefficient that has been tuned through comprehensive readability experiments [4].

Web pages contain text strings that do not belong to the information content of pages, but are instead used to format, structure and layout (e.g., tags) and to embed functionalities (e.g., scripts). The presence of these strings affect both word length and sentence length. While the use of HTML parsers allow to remove all strings not associated with the actual informative content of the pages (and thus reducing the errors in estimating word lengths), the estimation of sentence lengths is heavily affected by how the text is extracted from web pages, as we show with a concrete example in Section 3. This is because web pages are rich in tables, menus, lists, figures, captions, titles and subtitles: these are often part of the information content of the pages, but do not follow the expected structure of a sentence as assumed by the traditional readability measures. For example, often titles, menus and lists do not end with a punctuation mark that delimits the end of the sentence. In this paper we determine the effects that different ways of pre-processing web pages to extract the text associated with their information content have on the estimation of readability scores.

It is interesting to note that, already in the 1960s, the precise identification of sentence boundaries was a topic of concern for evaluating the readability of text. For example, Smith and Senter [14], authors of the Automated Readability Index (ARI), recommend typists to add to the end of each sentence an equal sign, aligned with a full stop, so as to explicitly demarcate sentence boundaries.

3. PRE-PROCESSING OF WEB PAGES

Before it is possible to estimate the readability of web pages, it is required to remove the HTML tags and the boilerplate text, so as to maintain only the text associated with the information content of the web pages. One approach to perform this text cleansing process is to use standard HTML parsing tools such as JSoup[1] for Java or Beautiful Soup[2] for Python. We used Beautiful Soup version 4.3.2, and we term this approach as **Naïve**.

Table 1: Five of the most used readability measures. W is the number of words in the text, Sy the number of syllables, S the number of sentences, C the number of characters, PW the number of polysyllables words (words with more than 3 syllables).

Automated Readability Index (ARI) [14]
$\text{ARI} = 4.71 \times \frac{C}{W} + 0.5 \times \frac{W}{S} - 21.43$
Coleman-Liau Index (CLI) [2]
$\text{CLI} = 5.89 \times \frac{C}{W} - (30.0 \times \frac{S}{W}) - 15.8$
Flesch-Kincaid Grade Level (FKGL) [10]
$\text{FKGL} = 0.39 \times \frac{W}{S} + 11.8 \times \frac{Sy}{W} - 15.59$
Gunning Fog Index (GFI) [8]
$\text{GFI} = 0.4 * (\frac{W}{S} + 100.0 \times \frac{PW}{W})$
Simple Measure of Gobbledygook (SMOG) [12]
$\text{SMOG} = 1.0430 * \sqrt{PW \times \frac{30.0}{S}} + 3.1291$

We also consider two open sourced tools developed specifically for removing the boilerplate from HTML pages: **Boilerpipe** [11] and **Justext** [13]. We used the Python version 1.2.0.0 of Boilerpipe[3] and version 2.1.1 of Justext[4].

Figure 1 shows the output of the *Naïve* approach applied to the fist paragraph of a web page from Wikipedia. The extracted text often presents an interesting characteristic: it lacks the punctuation marks to delimit the sentence boundaries. This has a clear effect on the readability measures that consider sentence length as an indication of text difficulty. To better understand the effect of this, we explore two possible approaches:

1. a sentence boundary (full stop) is added at the end of a line if no punctuation mark is found, resulting in **Short** sentences;
2. no sentence boundary is added, possibly resulting in **Long** sentences.

Note that the *Naïve-Short* approach is often used when processing web pages to automatically estimate readability measures, see for example [18].

4. READABILITY EVALUATION

To better understand the impact that parsing methods and approaches to sentence boundaries have when processing web pages, we consider the CLEF 2014 eHealth Task 3 collection, along with the 50 topics used in 2014 [6]. This collection contains web pages related to the medical domain and is used as a resource to evaluate search engines tailored to health consumers. We use this collection because of the importance the readability (and, more generally, the understandability) of web pages presenting medical advice has within consumer health search [16, 18].

[1] http://jsoup.org
[2] https://pypi.python.org/pypi/beautifulsoup4
[3] https://pypi.python.org/pypi/boilerpipe
[4] https://pypi.python.org/pypi/jusText

```
<body class="mediawiki page-Readability skin-vector action-view">
        <div id="siteNotice"><!-- CentralNotice --></div>
        <h1 id="firstHeading" class="firstHeading" lang="en">Readability</h1>
        <div id="siteSub">From Wikipedia, the free encyclopedia</div>
        <div id="jump-to-nav" class="mw-jump"> Jump to:  <a href="#mw-head">navigation</a>, <a href="#p-search">search</a> </div>
        <div id="mw-content-text" lang="en" dir="ltr" class="mw-content-ltr">
            <p>Readability</b> is the ease with which a text can be understood.
        </div>
</body>
```

CentralNotice	CentralNotice.
Readability	Readability.
From Wikipedia, the free encyclopedia	From Wikipedia, the free encyclopedia.
Jump to: navigation, search	Jump to: navigation, search.
Readability is the ease with which a text can be understood.	Readability is the ease with which a text can be understood.

Figure 1: Simplified Wikipedia entry for Readability (top) and the output of *Naïve* (bottom). In the bottom part, we show the result of the pre-processing approach termed *Long* (left), which does not modify the text extracted by the HTML parser, and that of the alternative pre-processing approach termed *Short* (right), which adds a full stop as sentence boundary at the end of every line. The *Long* approach concatenates all the text till it reaches a sentence boundary, producing longer sentences than the *Short* one.

Table 2: Number of words and sentences (mean and standard deviation) for documents CLEF 2014 eHealth Task 3 collection, as obtained by the three pre-processing tools and the two approaches to sentence boundaries (see Section 3).

Tool	# Words	# Sentences	
		Short	Long
Naïve	1001.5 ± 2062	137.2 ± 443	37.9 ± 93
Boilerpipe	364.2 ± 884	24.4 ± 55	18.6 ± 49
Justext	409.9 ± 1403	24.4 ± 82	19.4 ± 68

Table 2 reports the average number of words and sentences in the CLEF 2014 collection as extracted by the different pre-processing methods. These statistics are at the basis of the readability measures of Table 1. From the table, we can observe that the *Naïve* method produces one order of magnitude more words and sentences (except when using *Long*) than the other two methods. While small, the differences between *Boilerpipe* and *Justext* are still significant in that they can influence the estimations of readability measures. Similarly, the use of the *Short* approach for sentence boundary rather than the *Long* produces significant differences among all text pre-processing approaches.

We use the default vector space retrieval model of Apache Lucene 4.8 to retrieve the top 1,000 documents per query, using the query titles of the topics in the CLEF 2014 collection. For each query, we compute the readability scores of the retrieved documents according to the different settings considered here in terms of pre-processing tools and the approaches to sentence boundaries.

Figure 2 reports the mean values of readability scores averaged over the 50 topics for each combination of pre-processing and sentence boundary approaches. The results suggest that the choice of approach to use for sentence boundary has a significant influence on readability measures: the variance between the readability scores obtained with *Long* and *Short* is large across all methods, apart for CLI that appears to be the most robust readability formula in this aspect. For example, when using the *Naïve* pre-processing method, the mean readability score of ARI can vary more than 100%, from 10.9 ± 0.4, when using *Short* sentences to 22.5 ± 1.8, when using *Long* sentences. This high variability in the estimation of readability measures influences the conclusions one would infer about the difficulty of the retrieved documents: pages that could be readable by high school students (grade 11) – when sentence boundaries are detected with *Short* – become suddenly intractable for people with level of education below that of a PhD student (grade 22) – according to the readability measures computed using the *Long* method. In addition, note that different pre-processing methods (i.e., *Naïve*, *Boilerpipe*, *Justext*) lead to different conclusions about the readability of text. For example, Figure 2 suggests that, when ARI is used as readability measure and the *Long* approach is employed to identify sentence boundaries, *Naïve* and *Justext* provide contrasting results, with the mean readability of text assessed as being 22.5 ± 1.8 according to *Naïve* and 14.1 ± 0.6 according to *Justext*. These results highlight the significance that choices of pre-processing tool and sentence boundary identification approach have on the estimation of readability scores for web pages, when using the commonly adopted readability measures considered in this study.

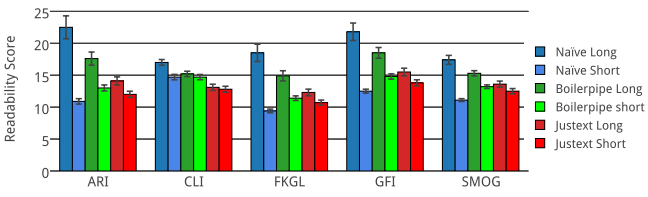

Figure 2: Readability scores for each measure based on different pre-processing and sentence boundary methods. Error bars indicate 95% confidence intervals around the mean.

The results in Figure 2 also suggest that CLI is the most stable readability measure among those considered in this paper. In particular, variations in pre-processing tool and sentence boundary identification have little impact on the estimated readability scores for this measure. The stability of CLI is due to the fact that $W \gg S$ and thus $\frac{S}{W} < 0$, dampening the effect of the relation between the number of words and sentences (in our experiments, $1 < 30.0 \times \frac{S}{W} < 4$), and ensuring stability across different values of S. This is unlike measures such as ARI, where $3 < 0.5 \times \frac{W}{S} < 13$.

Next, we consider how similar document rankings obtained from readability measure estimations are when using different pre-processing and sentence boundary approaches. This is interesting for information retrieval because it is often these differences between rankings, rather than the actual absolute value of the readability estimation, that are used to demote or promote web pages when taking into ac-

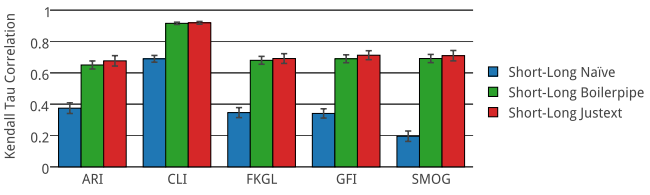

Figure 3: Kendall τ correlation and 95% confidence intervals between the *Short* and *Long* approaches for sentence boundary identification.

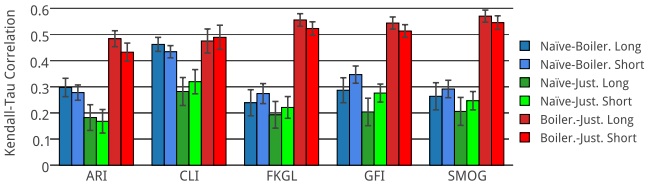

Figure 4: Kendall τ correlation and 95% confidence intervals between the approaches for HTML pre-processing, under different settings for sentence boundary identification.

count reading levels. For example, a high correlation between two different pre-processing settings would suggest that, although the actual readability scores may be very different, the preference ordering obtained by the readability measures (i.e., the ranking according to readability scores) are similar and therefore these two pre-processing settings would lead to little difference in terms of impact on retrieval.

To this aim, we consider the Kendall τ ranking correlation between different settings of sentence boundary identification (Figure 3) and pre-processing tool (Figure 4).

Figure 3 shows that, independently of the pre-processing tool used, the correlation between *Long* and *Short* rankings obtained when using *Boilerpipe* or *Justext* as pre-processing tools is generally high, with the maximum achieved using CLI ($\tau = 0.92 \pm 0.01$). However, if the *Naïve* approach to text pre-processing is used, then correlations deteriorate, with the SMOG measure exhibiting only marginal correlation ($\tau = 0.20 \pm 0.03$). CLI exhibits the least variance in correlation among the three pre-processing approaches (and indeed, the highest correlations) – a stability that was already observed when analysing Figure 2.

The results of Figure 4 suggest that different pre-processing tools produce different document rankings (when using readability to rank). Specifically, the highest correlation between two of these tools is achieved by the *Boilerpipe-Justext* pair – but these exhibit correlations of *only* about 0.5, independently of the readability measure or the sentence boundary approach (the highest correlation is achieved for SMOG by *Boilerpipe-Justext* using *Long*: $\tau = 0.57 \pm 0.02$). When comparing these methods to the *Naïve* approach, correlation sensibly decreases (apart for CLI that once again shows the smallest difference between settings).

5. CONCLUSION

This paper analysed the influence pre-processing and sentence boundary identification choices have on the estimation of readability measures for web pages. The experimental results show that these choices have a large impact on the estimation of readability scores, which in turn can drastically influence the order relations among documents that can be obtained from the readability scores. Our findings suggest that attention should be put on the choice of pre-processing settings when measuring readability for web pages. Advanced HTML cleansing tools such as *Boilerpipe* and *Justext* provide more stable results across settings. In addition, the use of the Coleman-Liau Index (CLI) as readability measure leads to the most stable results across choices of pre-processing tools and sentence boundary identification strategies (although we could not assess the quality of CLI for correctly estimating the readability of documents). In future work, we plan to study which combination of pre-processing settings and readability measure lead to estimations of readability that most agree with user assessments. The data and source code used in this work can be found online at: `https://github.com/joaopalotti/cikm_readability_2015`.

ACKNOWLEDGMENTS.

JP and AH are supported in part by the European Union Seventh Framework Programme (FP7/2007-2013) n°257528 (KHRESMOI), by Horizon 2020 program (H2020-ICT-2014-1) n°644753 (KCONNECT), and by the Austrian Science Fund (FWF) n°I1094-N23 (MUCKE).

6. REFERENCES

[1] M. Baroni, F. Chantree, A. Kilgarriff, and S. Sharoff. Cleaneval: a competition for cleaning web pages. In *LREC*, 2008.

[2] M. Coleman and T. L. Liau. A Computer Readability Formula Designed for Machine Scoring. *JAP*, 1975.

[3] K. Collins-Thompson, P. N. Bennett, R. W. White, S. de la Chica, and D. Sontag. Personalizing Web Search Results by Reading Level. In *CIKM*, 2011.

[4] W. H. Dubay. The principles of readability. *Costa Mesa, CA: Impact Information*, 2004.

[5] S. Dumais. Putting searchers into search. In *SIGIR*, 2014.

[6] L. Goeuriot, L. Kelly, et al. Share/clef ehealth evaluation lab 2014, task 3: User-centred health information retrieval. In *Working Notes of CLEF*, 2014.

[7] A. Graesser, D. McNamara, M. Louwerse, and Z. Cai. Coh-Metrix: Analysis of Text on Cohesion and Language. *Behav Res Meth Instrum Comput*, 2004.

[8] R. Gunning. *The Technique of Clear Writing*. McGraw-Hill, 1952.

[9] A. Jatowt and K. Tanaka. Is wikipedia too difficult?: Comparative analysis of readability of wikipedia, simple wikipedia and britannica. In *CIKM*, 2012.

[10] J. Kincaid, R. Fishburne, R. Rogers, and B. Chissom. Derivation of New Readability Formulas for Navy Enlisted Personnel. Technical report, 1975.

[11] C. Kohlschütter, P. Fankhauser, and W. Nejdl. Boilerplate detection using shallow text features. In *WSDM*, 2010.

[12] G. H. McLaughlin. SMOG Grading - a New Readability Formula. *Journal of Reading*, 1969.

[13] J. Pomikálek. *Removing Boilerplate and Duplicate Content from Web Corpora*. PhD thesis, 2011.

[14] E. A. Smith and R. J. Senter. Automated Readability Index. Technical report, 1967.

[15] C. Tan, E. Gabrilovich, and B. Pang. To Each His Own: Personalized Content Selection based on Text Comprehensibility. In *WSDM*, 2012.

[16] R. C. Wiener and R. Wiener-Pla. Literacy, pregnancy and potential oral health changes: The internet and readability levels. *Maternal and child health journal*, 2013.

[17] X. Yan, D. Song, and X. Li. Concept-based document readability in domain specific information retrieval. In *CIKM'06*, 2006.

[18] G. Zuccon and B. Koopman. Integrating Understandability in the Evaluation of Consumer Health Search Engines. In *MedIR*, 2014.