Enhancing Blog Readability for Non-native English Readers in the Enterprise

Chen-Hsiang Yu

MIT CSAIL

32 Vassar St

Cambridge, MA 02139

chyu@mit.edu

Jennifer Thom-Santelli

IBM T.J. Watson Research

1 Rogers Street

Cambridge, MA 02142

jthomsa@us.ibm.com

David R. Millen

IBM T.J. Watson Research

1 Rogers Street

Cambridge, MA 02142

david_r_millen@us.ibm.com

Abstract

Blogs are an important platform for people to access and share information, particularly in corporate settings where users rely on these systems for their work. However, because a global enterprise is multilingual, not all employees can understand the shared information in these systems easily if the content is written in a user's non-native language. As a result, this research focuses on enhancing the readability of blogs in enterprise social software for this group of users. The pilot user study of Japanese and Chinese bloggers suggest there are two main challenges: finding an interesting blog post to read and encountering difficulties in reading blog posts as currently rendered. Based on these findings, we designed and implemented a Firefox extension, *Clearly*, which uses web customization techniques to improve these two levels of readability issues.

Keywords

Readability, blogs, enterprise social software

ACM Classification Keywords

H5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

General Terms

Design, Languages, Measurement

Introduction

Within the workplace, internal blogs help to encourage collaboration among diverse communities of readers by facilitating access to knowledge and resources [4]. However, because a global enterprise is multilingual, not all employees can understand the shared information in these systems easily if the content is written in a user's non-native language. As a result, this language barrier may discourage blog readers who are less familiar with the prevalent language of the organization; thus making it less likely for them to reap the full benefits of enterprise blogging. For this reason, we focus on enhancing the readability of enterprise blogs for non-native English readers.

Drawing from Carmel et al. [1], we observe that novice English readers develop browsing strategies, such as common sense recognition, for choosing and reading blog posts to cope with their relative lack of fluency. If the selected blog post is not in the reader's native language, he or she may turn to support strategies to augment their understanding of the text. For instance, readers can refer to dictionaries and focus on keywords and key terms to increase comprehension [6]. Finally, there are other factors that may influence blog readability generally that may impact a non-native English reader more severely, such as font legibility [3] and general usability issues (e.g. lack of whitespace, distractions through extraneous content, large chunks of text [7]).

In this paper, we present a novel reading application, *Clearly*, to utilize web customization techniques and leverage readers' geographic networks to improve blog readability in enterprise social software.

User Requirements

To understand user challenges in reading non-native language blogs, we interviewed 6 non-native English speakers (3 from China, 3 from Japan) who were active bloggers and readers of an enterprise blogging tool deployed within a large multinational technology organization. We followed a similar user-centered design process to [9] in which we conducted a formative, pre-design evaluation of needs of a sample of lead users. Regarding to the user sample, we used a sampling approach similar to [2] and [8]. The sample in our interview study is representative in terms of language ability; multilingual employees at this global corporation must demonstrate a certain level of proficiency in English in order to obtain employment.

On average, the 6 participants (5 M, 1 F) have been speaking and reading English for 17.2 (SD = 7.1) years. We conducted two of the interviews face-to-face, two over the phone and two through email. We used email in instances where participants preferred to have written questions to have more time to read and understand what was being asked of them. For this pilot study, we asked participants to browse and select two blog posts to read from the enterprise blogging tool and asked them a series of semi-structured interview questions during these tasks.

In sum, we observed two main challenges for non-native English readers: finding an interesting blog post to read and reading blog posts as currently rendered. The first challenge relates to the initial list display of the blog posts, which participants found difficult to scan and filter, mainly because of the number of items shown. Participants did not use the current filtering features (e.g. tag cloud, featured blogs, most

commented blogs) except for a filter indicating blog posts recommended by other users, but relied mainly on scanning. The 3 Japanese participants also chose to read blog posts that were in their native language, when available.

The second challenge deals with the blog posts itself, once one has been selected for reading. Participants commented that they did not attend to anything in the left hand navigation or the right sidebar while reading the blog post. In addition, participants expressed interest in features that would enable support strategies that would help them make sense of a blog post [6]. For instance, they requested resources, such as translation of terms into their native languages and highlighting of topic keywords.

System Design and Implementation

The context in which readability is applied in this study differs slightly from prior research describing readability for non-native English readers [9]. Such work focused on the display of longer blocks of text and stripping out content that serves as distractions, which are strategies employed by other similar systems (e.g. Readability, Safari 5's Reader). These techniques remain useful for formatting blog posts, once selected by the reader. We observe, however, that finding a blog post of interest by reading and scanning a long list of entries also poses difficulty for non-native English speakers. As a result, we include the initial list display, as well as the blog post itself, as interfaces that can be enhanced for readability.

Clearly: A Prototype to Enhance Blog Readability

We then implemented a Firefox extension, *Clearly*, to address the readability issues in both the blog list and blog post content view. *Clearly* uses a knowledge repository to assist in reading by providing resources for support strategies (e,g, reference tools, highlighted text). The knowledge repository is a JSON-format structure created dynamically and includes a multilingual vocabulary dictionary and a terminology book to define enterprise-specific words.
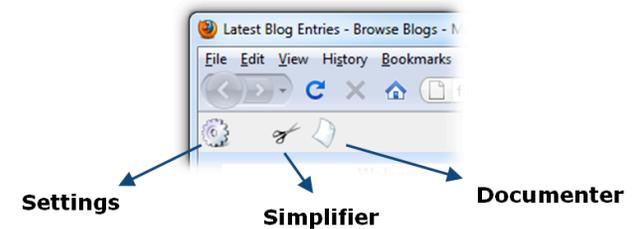

Figure 1. The *Clearly* toolbar.

As Figure 1 illustrates, there are only three buttons in the design: Settings, Simplifier and Documenter. When the user clicks on the Settings button, a sidebar shows up to collect the reader's personal information (e.g e-mail, preferred language, keywords of interest) to construct the user's Reading Settings. Inspired by [5], *Clearly* uses geographic filter to support participants' desire to find content of nearby co-workers and to improve discovery of interesting content. We use the reader's email address as a key to make a call to the corporate directory's API to obtain employee location. This geographic information is used to construct a three-level relationship: close-to-reader, the-same-country, different-country. We save the indicated

preferences to a local database; in this case, the Firefox browser's default database, SQLite, for local storage.
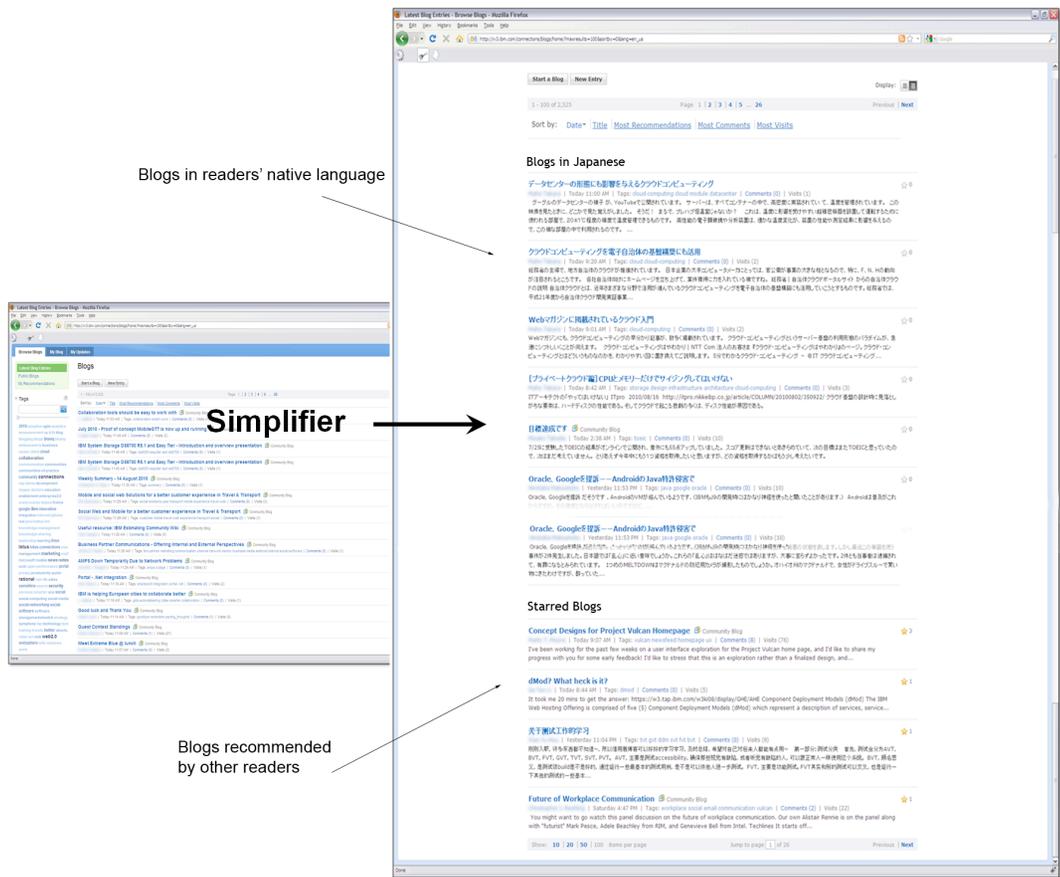

Figure 2.The Simplifier transformation with native language and reader-recommended blog filters.

As Figure 2 illustrates, when the reader clicks on the Simplifier button, the application uses JavaScript with rule-based methods to parse the DOM tree of the currently viewed HTML page and extract blog post entries for content filtering into the three categories gathered from the Reader Settings: reader's native language, blog posts recommended by other readers, and blog authors whose geographic location is close to the reader. We detect the language of each blog post by sending random words from the blog title and abstract to a web-based language detector (e.g. Google Language Detector).

This filtering addresses the concerns of the first level of blog readability in aiding the scanning and browsing activities of the non-native English reader. The Simplifier button decreases distraction by transforming the right-hand and left-hand columns to increase whitespace. To provide additional context for recognition and topic selection for the novice browser, we increased the number of words in the snippet text. We also addressed the user-expressed preference for native-language blogs by emphasizing those posts in its own category. Filtering blog posts by geography serves two purposes: to facilitate discovery of nearby co-workers who are not related in an org-chart explicitly and to uncover blog posts from writers whose native language may be the same as the reader's.

When a reader chooses a blog post to read, he or she can click on the Documenter button to apply the content transformations that address the second level of blog readability to provide resources for the support strategies used by non-native English readers (e.g. topic keyword highlighting and translation) [6]. As shown in Figure 3, there are two distinct formatting

highlighted topic keywords and line spacing with vocabulary translation. We provide a slider for the user to switch between these two transformations.

When the Documenter button is selected, the original blog content is transformed to the first format, which improves the general usability issues by removing the content from the left-hand and right-hand columns to decrease distractions and adding additional whitespace between paragraphs for increased legibility. In addition, we change font type and line width and highlight topic keywords that have been previously selected by the reader. When the reader switches the slider to line spacing with translation format, the line-height of the viewing page will become three times larger to give more whitespace between lines.

From the knowledge repository created during *Clearly's* initialization process, we assemble a vocabulary set containing English words of intermediate and advanced difficulty drawn from shared publicly available English-learning resources. When a user initiates the Documenter function, the application tokenizes the blog post content. If any terms from the selected blog post match one in the vocabulary set, we provide translation for each term in the reader's native language. Each of the translated words is highlighted, but we also provide a close button to each annotated word to remove the translation support if desired. As the reader uses the Documenter feature in *Clearly* to read multiple blog posts or becomes more confident in their knowledge of terminology in these blogs, we save these preferences so that users do not encounter redundant translation.
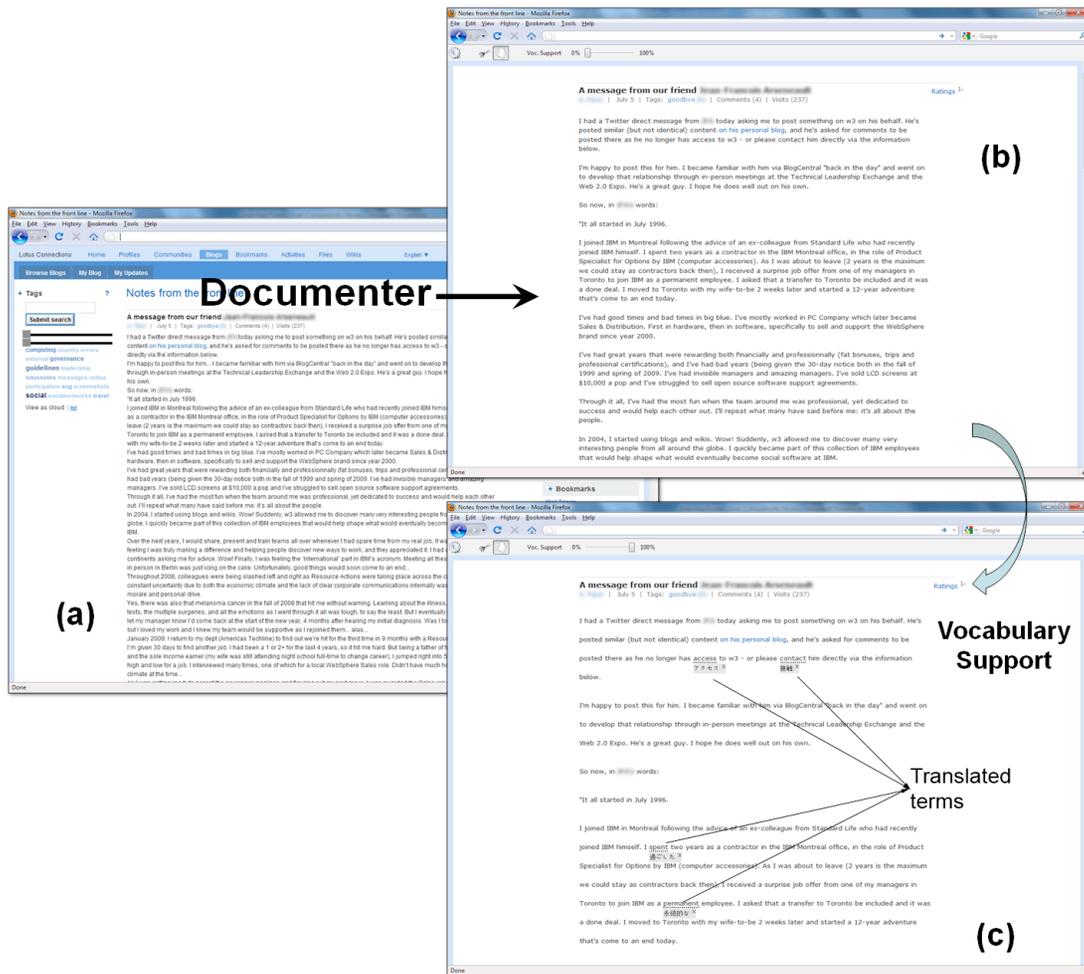

Figure 3. The Documenter with blog page transformation and vocabulary support. (a) Original blog post (b) Advanced content presentation with highlighted keywords. (c) Line spacing with translated terms.

transformations: advanced content presentation with

In summary, *Clearly* provides multiple levels of support to non-native English blog readers through two main transformations. First, the Simplifier aids readers in the scanning and browsing process by filtering content into native language, reader-recommended, and geographic proximity filters. Second, the Documenter augments the strategies employed by non-native English readers by highlighting topic keywords and providing translation of difficult vocabulary.

Conclusion and Future Work

Our motivation for this work draws from our observation that non-native language readers may not avail themselves of the potential resources found within enterprise blogs. Based on research describing the browsing behaviors of novices and the support strategies employed by English-as-a-Second Language readers, we designed *Clearly*, a novel application that enhances blog readability on two levels: by improving the browse and scan process and providing additional resources for readers to use support strategies if desired. We attempt to improve online readability in a social software environment by adopting two approaches: (1) providing support for finding interesting and relevant content to read and (2) providing support for reading the selected text. Future work includes large-scale deployment and evaluation with non-native English readers in the organization.

References

[1] Carmel, E., Crawford, S., and Chen, H. Browsing in hypertext: A cognitive study. *IEEE Transactions on Systems, Man and Cybernetics 22*, 5 (1992), 865-884.

[2] DiMicco, JM, Millen, D., Geyer, W., Dugan, C. Brownholtz, B. and Muller, M. Motivations for Social Networking at Work. *Proc.CSCW2008*, ACM (2008).

[3] Dillon, A., Kleinman, L., Choi, G.O., and Bias, R. Visual search and reading tasks using ClearType and regular displays: two experiments. *Proc.CHI2006*, ACM (2006).

[4] Huh, J., Jones, L., Erickson, T., Kellogg, W.A., Bellamy, R.K.E., and Thomas, J.C. BlogCentral: the role of internal blogs at work. *Ext.Abstracts CHI2007*, ACM (2007).

[5] Liben-Nowell, D., Novak, J., Kumar, R., Raghavanm P. and Tomkins, A. Geographic Routing in Social Networks. *Proceedings of the National Academy of Sciences*, Volume 102, Number 33, p.11623-11628, (2005).

[6] Sheorey, R. and Mokhtari, K. Differences in the metacognitive awareness of reading strategies among native and non-native readers. *System 29*, 4 (2001), 431-449.

[7] Spool, J.M., Schroeder, W., Scanlon, T., and Snyder, C. Web sites that work: Designing with your eyes open. *Proc.CHI1998*, ACM (1998).

[8] Thom-Santelli, J. and Millen, D. Learning by Seeing: Photo Viewing in the Workplace. *Proc.CHI2009*, ACM (2009).

[9] Yu, C.H. and Miller, R.C. Enhancing web page readability for non-native readers. *Proc.CHI2010*, ACM (2010).