ConsentCanvas: Automatic Texturing for Improved Readability in End-User License Agreements

Oliver Schneider & Alex Garnett
Department of Computer Science, University of British Columbia
201-2366 Main Mall, Vancouver, BC, Canada, V6T 1Z4
oschneid@cs.ubc.ca, axfelix@gmail.com

Abstract

We present ConsentCanvas, a system which structures and "texturizes" End-User License Agreement (EULA) documents to be more readable. The system aims to help users better understand the terms under which they are providing their informed consent. ConsentCanvas receives unstructured text documents as input and uses unsupervised natural language processing methods to embellish the source document using a linked stylesheet. Unlike similar usable security projects which employ summarization techniques, our system preserves the contents of the source document, minimizing the cognitive and legal burden for both the end user and the licensor. Our system does not require a corpus for training.

1 Introduction

Less than 2% of users read End-User License Agreement (EULA) documents when indicating their consent to the software installation process (Good et al., 2007). While these documents often serve as a user's sole direct interaction with the legal terms of the software, they are usually not read, as they are presented in such a way as is divorced from the use of the software itself (Friedman et al., 2005). To address this, Kay and Terry (2010) developed what they call *Textured Consent* agreements which employ a linked stylesheet to augment salient parts of a EULA document. Unlike summarization-driven approaches to usable security, this is achieved without any modification of the underlying text, minimizing the cognitive and legal burden for both the end user and the licensor and removing the need to make available a supplementary unmodified document (Kelley et al, 2009; Farzindar, 2004).

We have developed a system, ConsentCanvas, for automating the creation of a Textured Consent document from an unstructured EULA based on the example XHTML/CSS template provided by Kay and Terry (2010; Figure 1). Our system does not currently use any complex syntactic or semantic information from the source document. Instead, it makes use of regular expressions and correlation functions to identify variable-length relevant phrases (Kim and Chan, 2004) to alter the document's structure and appearance.

We report on ConsentCanvas as a work in progress. The system automates the labour intensive manual process used by Kay and Terry (2010). ConsentCanvas has a working implementation, but has not yet been formally evaluated. We also present the first available implementation of Kim and Chan's algorithm (2004).

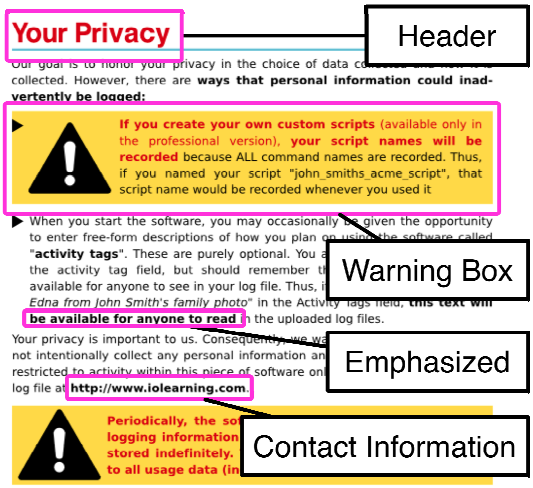

Figure 1. Example Textured Consent Document as designed by Kay and Terry (2010).

Proceedings of the ACL-HLT 2011 Student Session, pages 41–45,
Portland, OR, USA 19-24 June 2011.

2 Methods

We built ConsentCanvas in Python 2.6 using the Natural Language Toolkit (NLTK) 2.0b9. It uses a modified version of the markup.py library available from http://markup.sourceforge.net to generate valid HTML5 documents. A detailed specification of our system workflow is provided in Figure 2. ConsentCanvas was designed with modularity as a priority in order to adapt to the needs of future experimentation and improvement. As such, we contribute not just a working application, but also an extensible framework for the visual embellishment of plaintext documents.

2.1 Analysis

Our system takes plain-text EULA documents as input through a simple command line interface. It then passes this document to four independent submodules for analysis. Each submodule stores the initial and final character positions of a string selected from within the document body, but does not modify the document before reaching the renderer step. This allows for easy extensibility of the system

2.2 Variable-Length Phrase Finder

The variable-length phrase finder module features a Python implementation of the Variable-Length Phrase Finding (VLPF) Algorithm by Kim and Chan (2004). Kim and Chan's algorithm was chosen for its domain independence and adaptability, as it can be fine-tuned to use different correlation functions.

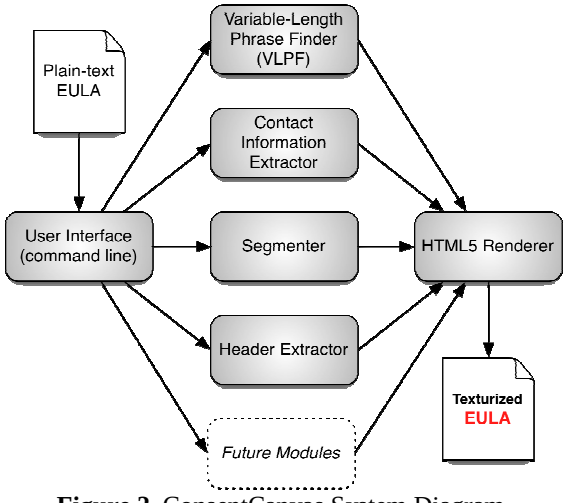

Figure 2. ConsentCanvas System Diagram.

This algorithm computes the conditional probability for the relative importance of variable-length n-gram phrases from the source document alone. It begins by considering every word a phrase with a length of one. The algorithm iteratively increases the length of phrases, adding an adjacent word to the end. That is, every phrase of length m P{m} is considered as P{m-1}w, where w is a following adjacent word.

Correlation is calculated between the leading phrase P{m-1} and the trailing word w. Phrases that maintain a high level of correlation are creating by appending the trailing word w, and those with a correlation score below a certain threshold are pruned before the next iteration. This continues until no more phrases can be created. This method is completely unsupervised.

The VLPF algorithm is able to use any of several existing correlation functions. We have implemented the Piatetsky-Shapiro correlation function, the simplest of the three best-performing functions used by Kim and Chan, which achieved a correlation of 92.0% with human rankings of meaningful phrases (2004).

We removed English stopwords, but did not perform any stemming when selecting relevant phrases because the selection of VLPs did not depend on global term co-occurrence, and we did not want to modify selected exact phrases. We emphasize the top 15% meaningful phrases (as determined by the algorithm) for the entire document. 15% was chosen for its comparable results to Kay and Terry's example document (2010). The phrase selected as the most relevant is also reproduced in the pull quote at the top of the document, as shown in Figure 3.

2.3 Contact Information Extractor

The contact information extractor module uses regular expressions to match URLs, email addresses, or phone numbers within the document text. This information was displayed as bold type in accordance with the Textured Consent template.

2.4 Segmenter

The segmenter module uses Hearst's TextTiling algorithm to "segment text into multi-paragraph subtopic passages" (1997). This algorithm analyzes

patterns of lexical co-occurrence and distribution in order to impose topic boundaries on a document. ConsentCanvas uses the NLTK implementation of the TextTiling algorithm. Segmentation was not applied to the entire document (doing this resulted in a messy layout incoherent with structuring applied by headers and titles). Instead, we used it to identify the lead paragraph of the document, which was rendered differently using the "lead paragraph" container in the template. Future versions will use a more modern segmenting algorithm.

2.5 Header Extractor

The header extractor module uses regular expressions to match any section header-like text from the original document. Several different search strings were used to catch multiple potential header types, including but not limited to:

- 8 OR FEWER ALL-CAPS TOKENS
- 3. Single level numbered headers
- 3.1 Multi-level numbered headers
- Eight or fewer tokens separated by a line break

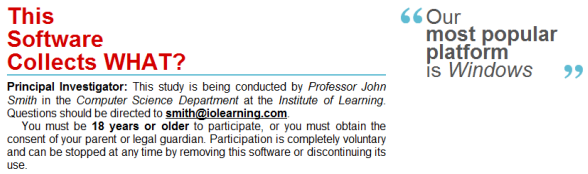

Figure 3. Summary text in the example document.

2.6 Rendering

Each analysis submodule produces a list of character positions where found items begin and end. These are passed to our rendering system, which inserts the corresponding HTML5 tags at the positions in original plaintext EULA. We append a header to the output document to include the linked stylesheet per HTML5 specifications.

3 Analysis & Results

We conducted a brief qualitative analysis on ConsentCanvas after implementation and debugging. However, the problem space and system are not yet ready for formal verification or experimentation. More exploration and refinement are required before we will be able to empirically determine if we have improved readability and comprehension.

3.1 Corpus

We conducted our analysis on a small sample of EULAs from the same collection used by Lavesson et al. (2008) in their work on the classification of EULAs. There were 1021 EULAs in this corpus divided into 96 "bad" and 925 "good" examples. We used the "good" examples for our analysis.

3.2 Variable-Length Phrase Finding Results

Variable-Length Phrases (VLPs) were reasonably effective. In several of the best examples of texturized EULAs security concerns were highlighted; in the texturized version of one document, the pull quote was "on media, ICONIX, Inc. warrants that such media is free from defects in materials and workmanship under normal use for a period of ninety (90) days from the date of purchase as evidenced by a copy of the receipt. ICONIX, Inc. warrants." In the same EULA, other VLPs proved helpful: "e that ICONIX, Inc. is free to use any ideas, concepts," "(except one copy for backup purposes)," and "Inc. ICONIX, Inc. does not collect any personally identifiable information regarding senders." Some phrases have incomplete words at the beginning and end; this is an artifact of a known but unfixed bug in the implementation, not a result of the algorithm.

However, these results were mixed in other EULAs. Several short but frequent phrases were found to be VLPs, such as "Inc.," in the same EULA. In short licenses consisting of only one to three paragraphs, sometimes no relevant VLPs were discovered. There are also many phrases that should be highlighted that are not.

3.3 Preliminary System Evaluation

We conducted an informal evaluation in which our system applied texture to 15 documents chosen from our corpus at random. Of these, five were determined to be highly readable exemplar documents. An excerpt from one of these is shown in Figure 4. Of the remaining ten documents, four had poorly selected header markup but were otherwise satisfactory, two were too short or poorly-structured to benefit from the insertion of header markup, two did not perform well on the VLPF step, and two had several errors which appeared to have been caused by the use of non-ASCII characters in the original document.

The pull quote text was nearly unintelligible in almost all cases, due largely to the fact that it did not split evenly on sentence borders. We did not let this detract from our evaluation of the documents, because performance in this area was so consistently, and charmingly, poor, but did not affect readability of the main document body.

4 Discussion

Our preliminary analysis has provided several insights into the challenges and next steps in accomplishing this task.

4.1 Comparisons with Kay and Terry

Kay and Terry (2010) make reference to "augmenting and embellishing" the document text – specifically *not* altering the original content. However, their example document is written concisely in a user-friendly voice dissimilar to most formal EULAs found in the wild. Their work provides a strong proof of concept, but a key line of investigation will be whether their approach is practical, or whether some preprocessing is necessary to simplify content.

4.2 Handling Legal Language

We had anticipated a considerable amount of difficulty in selecting meaningful phrases from difficult-to-understand legal language in the source document. However, most documents were found to contain a number of high-frequency VLPs with both layperson-salient legal terminology and common clues to document structure.

4.3 Future Work

ConsentCanvas is fully implemented but offers many opportunities for improvement as the task becomes better understood. The variable-length phrase finding module only incorporates a single correlation function. More will be added, drawing in particular from those documented by Kim and Chan (2004). Machine learning techniques might also be used to classify phrases as relevant or not, leading to better-emphasized content.

The *rhythm* of emphasized phrasing is also important. In the example license designed by Kay and Terry (2010), there are one or two emphasized phrases in each section. The phrases found by ConsentCanvas are often sporadic, clustering in some sections and absent from others. As a result of this, readability suffers, and so we may need to look into possible stratification of VLPs. This might also aid multi-lingual documents, of which there are a few examples (a cursory look showed the results in French were comparable to those in English in a bilingual EULA in our corpus).

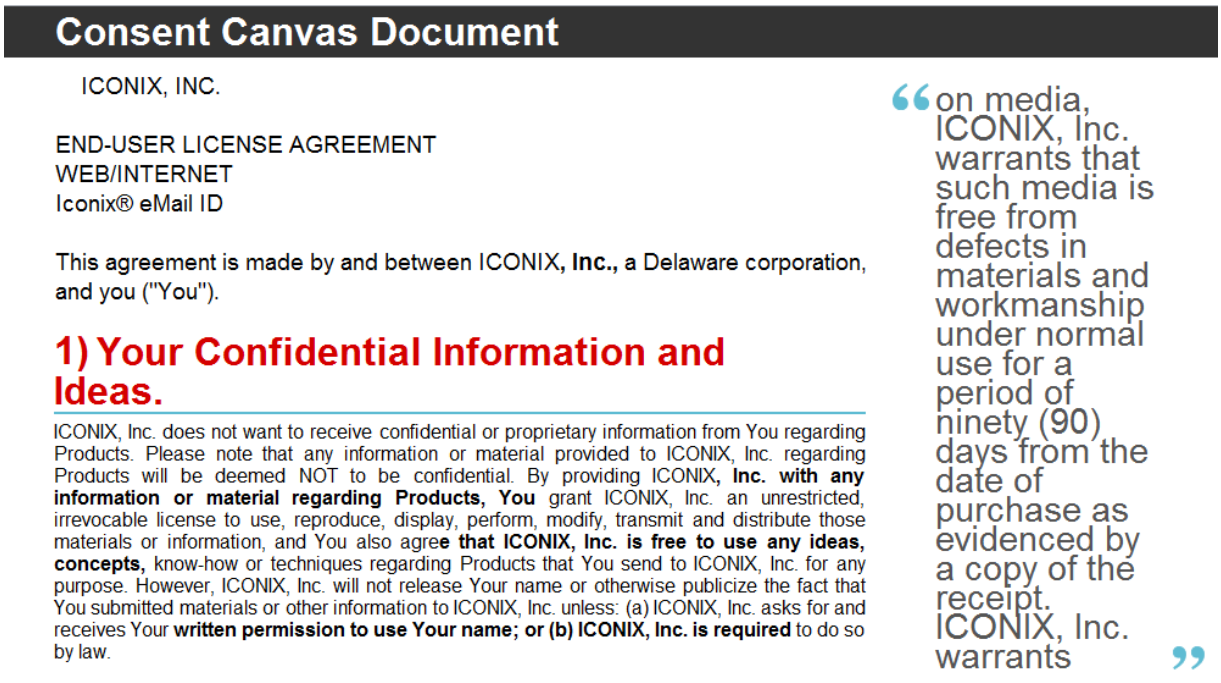

Figure 4. Summary text in an example output document.

Contact information is currently emphasized in the same manner as salient phrases. We plan to eventually embed hyperlinks for all URLs and email addresses found in the source document, as in Kay and Terry (2010).

The segmenter module uses the basic TextTiling algorithm with default parameters. More recent approaches could be implemented and could act on more than the lead paragraph. For example, coherent sections of long EULAs might be identified and presented as separate containers.

We plan to improve header extractor providing more sophisticated regular expressions; we found that a wide variety of header styles were used. In particular, we plan to consider layouts that use digits, punctuation, or inconsistent capitalization in multiple instances in the document body.

There is currently no module that incorporates the "Warning" box from Kay and Terry (2010). This module would be designed to select relevant multiline blocks of text by using techniques similar to the variable-length phrase finder or the segmenter.

ConsentCanvas will also be extended to support command-line parameters. This will enable customized texturing of EULAs and facilitate experimentation for understanding and evaluating gains in comprehension and readability. Finally, we will conduct a formal user evaluation of ConsentCanvas.

5 Conclusion

We have provided a description of the work in progress for ConsentCanvas, a system for automatically adding texture to EULAs to improve readability and comprehension. Informal analysis revealed several key challenges in accomplishing this task and identified the next steps towards exploring effective solutions to this problem.

Acknowledgments

We would like to thank the reviewers for their helpful feedback and Dr. Giuseppe Carenini for his support and encouragement. This work was partially supported by an NSERC CGS M scholarship.

Appendix

The source code, our corpus, and a sample of converted documents are all available at: https://github.com/axfelix/consentCanvas.

References

Farzindar, A. 2004. Legal text summarization by exploration of the thematic structures and argumentative roles. *Text Summarization Branches Out.*

Friedman, B. 2005. Informed consent by design. In *Security and Usability*, Eds. Lorrie Faith Cranor & Simson Garfinkel,

Good, N., Dhamija, R., Grossklags, J., Thaw, D., Aronowitz, S., Mulligan, D. and Konstan, J. 2005. Stopping spyware at the gate: a user study of privacy, notice and spyware. *Proceedings of the 1st Symposium on Usable Privacy and Security*. 43–52.

Hearst, M.A. 1997. TextTiling: segmenting text into multi-paragraph subtopic passages. *Computational linguistics 23*, 1: 33–64.

Kay, M. and Terry, M. 2010. Textured agreements: Re-envisioning electronic consent. *Proceedings of the Sixth Symposium on Usable Privacy and Security.*

Kelley, P.G., Bresee, J., Cranor, L.F., and Reeder, R.W. 2009. A nutrition label for privacy. *Proceedings of the 5th Symposium on Usable Privacy and Security*: 1–12.

Kim, H. and Chan, P.K. 2004. Identifying variable-length meaningful phrases with correlation functions. *16th IEEE International Conference on Tools with Artificial Intelligence*, 30-38.

Lavesson, N., Davidsson, P., Boldt, M., Jacobsson, A. 2008. Spyware Prevention by Classifying End User License Agreements. *Studies in Computational Intelligence, volume 134.* 373-382.