Supervised and Unsupervised Methods for Robust Separation of Section Titles and Prose Text in Web Documents

Abhijith Athreya Mysore Gopinath[1], Shomir Wilson[1] and Norman Sadeh[2]
[1]College of IST, Pennsylvania State University, University Park, PA 16802, USA
[2]School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA
abhijith@psu.edu, shomir@psu.edu, sadeh@cs.cmu.edu

Abstract

The text in many web documents is organized into a hierarchy of section titles and corresponding prose content, a structure which provides potentially exploitable information on discourse structure and topicality. However, this organization is generally discarded during text collection, and collecting it is not straightforward: the same visual organization can be implemented in a myriad of different ways in the underlying HTML. To remedy this, we present a flexible system for automatically extracting the hierarchical section titles and prose organization of web documents irrespective of differences in HTML representation. This system uses features from syntax, semantics, discourse and markup to build two models which classify HTML text into section titles and prose text. When tested on three different domains of web text, our domain-independent system achieves an overall precision of 0.82 and a recall of 0.98. The domain-dependent variation produces very high precision (0.99) at the expense of recall (0.75). These results exhibit a robust level of accuracy suitable for enhancing question answering, information extraction, and summarization.[1]

1 Introduction

Web text continues to be an immense resource for researchers working in NLP and related areas, but its typographic structure (i.e., its visual organization) remains underutilized. Many texts on the web are organized into sections based on the topics presented, and each section has a title followed by prose text. The title tends to be visually distinct to separate it from the prose that succeeds it. Apart from improving the readability of the page, this explicit organization makes titles act as intuitive indexes for the prose text(s) that follow them. In other words, titles are concise summaries of the following prose text.

Most current methods of web text extraction do not separate titles from prose text, instead treating the entire document as a single unit without internal structure. However, this internal organization has the potential to provide input to a variety of NLP tasks that can make use of information on topicality or discourse structure. In the case of question answering, this information can facilitate identifying maximally relevant sections to search for answers. It can also increase accuracy by filtering out false positives and minor references related to the topic of the question, which might be present in other sections. For information retrieval tasks, semantic matching of the search terms can be performed on the titles first, followed by a search on the prose texts associated with closely matching headers.

However, detecting the titles and prose segments in an HTML document is difficult for two reasons. One of them is the flexibility of HTML, which allows the same typographic layout to be represented in code in multiple ways. Tags are also nested with varying depths. Figure 1 illustrates this problem: similar title and prose text segments from four website privacy policies[2] have altogether different HTML tag structures. The second problem is that it is not straightforward to distinguish the information (encoded in HTML) that is necessary for title-prose detection from the rest of the HTML structure, including unrelated links, multiple tags with little or no content and page headers and footers. Sieving only useful information from these pages requires a flexible approach.

[1]The source code and corpora generated by this research are available at https://github.com/abhijith-athreya/ASDUS.

[2]All the policies were retrieved on 2018-01-20, from the below URLs:
https://rule.alibaba.com/rule/detail/2034.htm
https://www.apple.com/legal/privacy/en-ww/
https://www.cbsinteractive.com/legal/cbsi/privacy-policy
https://help.bet365.com/en/privacy-policy

Visual Representation	HTML
A. COLLECTION OF INFORMATION 1. Your privacy is important to us and we have taken steps to ensure that we do not collect more information from you than is	<p>
Collection and Use of Personal Information Personal information is data that can be used to identify or contact a single perso	<section> <h5>
2. Information Collected **Information you provide directly to CBS Interactive Services.** You are not required t	<h4>
Information Collected and How It Is Used The information and data about you which we may collect, use and includes the following:	<b>

Figure 1: Excerpts from the website privacy policies of Alibaba.com, Apple, CBS Interactive and bet365. At left are browser renderings, and at right are the XPaths of the section titles.

We present ASDUS (Automatic Segment Detection using Unsupervised and Supervised learning), a system that uses a variety of features of text and markup structure to identify the title and prose organization of HTML documents automatically. Our approach effectively strips away all unrelated HTML and produces simplified HTML with a predictable tag structure, thus making the extraction of section titles and prose text straightforward for downstream applications. We present two approaches, a domain-independent approach that requires no prior training and a domain-dependent approach that takes advantage of a labeled corpus. The domain-independent approach (abbreviated *DI*) yields an overall precision of 0.82, recall of 0.98 and coverage of 97.21%. The domain dependent approach (*DD*) produces an overall precision of 0.99, recall of 0.75 and coverage of 93.10%. We are releasing ASDUS and associated datasets to the research community for use and improvement.

2 Related Work

The detection of titles and prose in web documents has received little (if any) prior attention, but we briefly survey literature in some related areas.

Titles can be thought of as metalinguistic descriptions for the prose text they are associated with. Wilson (2013) attempted to identify a core metalinguistic vocabulary for the English language and to automatically identify instances of metalanguage usage. Deixis present in a text can also be considered metadata and detection of deixis helps in structuring the flow of information. Wilson and Oberlander (2014) attempted to capture word senses related to deixis.

Topic classification is the problem of segregating a document into different topics, and argumentative zoning (Teufel et al., 1999) was an early effort that shares some goals with the present work, as it addressed the detection of the main thematic areas in research articles. Teufel and Kan (2011) built a robust argumentative zoning system which used maximum entropy modeling to go with morphological features. Conditional random fields were adopted for categorization of sentences of a scientific abstract into different sections by Hirohata et al. (2008). Using posterior discourse and lexical constraints as features, Guo et al. (2013) improved upon existing information structure analysis of scientific documents through unsupervised and minimal supervised learning.

HTML structure analysis is the process of extracting useful information by utilizing the underlying HTML document structure. Information extraction from HTML using machine learning was introduced in SRV (Freitag, 1998), a top-down relational algorithm for information extraction. This system aimed at filling pre-defined slots for a web page in a particular domain. A set of extraction rules suitable to extract information from a website is called a wrapper (Flesca et al., 2004). Dalvi et al. (2011) worked on enhancing wrapper induction techniques by introducing a generic framework which allows for training on noisy data. Liao et al. (2015) used web block segmentation and machine learning algorithms to retrieve business event data, such as coupons, tickets, and sales campaigns. García-Plaza et al. (2017) worked on using fuzzy logic to leverage HTML markup for web page clustering. Using four essential features viz., text frequency, title, emphasis and po-

sition, they define 31 independent rules to arrive at the importance of a text segment. Unlike our approach, which is independent of the tag structures and learns patterns on its own, these methods depend on handcrafted rules and similar tag structures to identify various sections of the document.

3 Approach

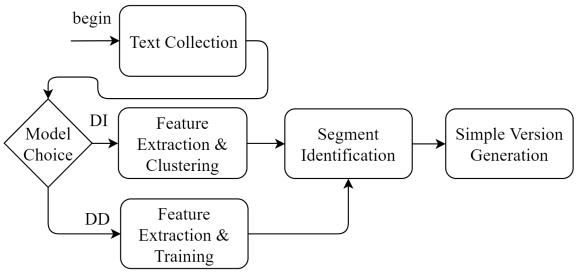

Figure 2: Stages of the DI and DD approaches.

3.1 Domain-Independent Approach (DI)

Figure 2 depicts the stages of this approach, which are explained in detail below.

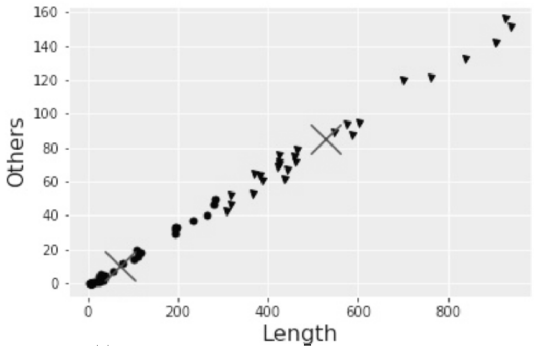

Figure 3: The graph shows two clusters. The one nearer to the origin is the title cluster and the farther one is the prose cluster. Cluster medians are marked with *X*s.

Text Collection: Using jsoup (Hedley, 2017), we parse the HTML file and for each non-empty tag encountered we extract a tuple consisting of the text and its XPath.

Feature Extraction and Clustering: For each tuple, we extract a set of features which aids in differentiating titles from prose text. The features are own text length, next text length, number of punctuation symbols, number of sentences, number of stop words (stop words were derived from Weka (Hall et al., 2009)), number of discourse cues, number of named entity slots and number of words with capitalized initial letter. We calculated the number of discourse cues as the sum of explicit discourse markers provided by Denver (2018) and the number of coreference chains. We used Stanford CoreNLP (Manning et al., 2014) for identifying coreference chains and named entities.

Titles tend to contain less text than prose segments, as well as fewer punctuation symbols, stop words, discourse markers, sentences and named entities. Since titles are not followed immediately by titles (in most cases), the next text length feature helps to remove false positives for title identification. All these features are collapsed into two dimensions, one being the text length and the other being the linear combination of remaining features.

K-means clustering (using scikit-learn (Pedregosa et al., 2011)) is performed on the feature set to group it into two mutually exclusive subsets. Figure 3 shows an example plot of the clustering. We obtain two distinct clusters with the title cluster closer to the origin and the prose cluster away from the origin. We leverage this property to obtain the label of the title cluster by identifying the label for the origin ([0,0]). Each text segment (and thus each tuple) is classified into title or prose using the k-means model.

Segment Identification: For each potential title identified in the previous phase, an overlap score is calculated by measuring the overlap between the lemmatized form of words belonging to the title and the lemmatized form of words from the next text segment. Based on experiments on a development set, the overlap threshold was set to 75%. Titles with overlap scores exceeding the threshold and their corresponding XPaths are added to the list of probable title candidates.

Simple Version Generation: Each element in the final title list is marked with a custom attribute in the original HTML. To generate the simple version, a top down parse of the HTML is performed, wherein nodes with custom attribute (titles) are enclosed in $<h2>$ tags and the text between two consecutive titles (prose) is enclosed in $<p>$ tags. For the last segment, the prose text immediately following the title is added. For subsequent prose sections, a unigram overlap score similar to the one in the previous step is performed to avoid the addition of unrelated text, such as a page footer. Legitimate textual content appearing before the first title is included in the final output under an uncategorized title.

Type of Doc.	# Docs.	# Titles	# Prose
Privacy	152	3611	14506
TOS	100	2299	7818
Misc.	51	685	3676

Table 1: Test set details of the DI approach.

3.2 Domain-Dependent Approach (DD)

The DD approach differs from the DI approach by feature selection: the DD approach trains a neural network classifier on previously labeled examples.

Feature Extraction and Training: To construct labeled data for training, we chose to use web privacy policies due to their lengthy nature, the presence of a reasonable number of segments, the presence of relatively similar content with varying HTML structure across websites and their importance to the general public. We annotated each HTML tag of 100 web privacy policies with one of three labels: title, prose and unrelated. We built two word embedding models (using gensim (Řehůřek and Sojka, 2010)), one using text from the titles and the other using prose texts. Then for each HTML tag, we calculated two semantic relatedness scores: one between the title embedding and the text in the tag (*t-score*), and the other between the prose embedding and the text in the tag (*p-score*). The t-score is the log probability for a text segment with respect to the title embedding. The p-score is the log probability for a text segment with respect to the prose embedding. The intuition behind using these scores is that the t-score will be higher for titles and lower for prose text. Similarly, the p-score will be higher for prose text than titles. The t-score, p-score and length of the text formed the feature set. Using Tensor Flow (Abadi et al., 2015), we trained a feed forward neural network (h1=24, h2=48) to classify text between tags as title, prose or unrelated.

4 Dataset and Results

The problem of automatic detection of titles and prose text in HTML documents has received scant prior attention. Due to this, a corpus containing HTML documents along with their respective annotated versions (titles and prose sections annotated) was unavailable. This lack of data prompted the creation of a new corpus consisting of web documents and their respective annotated versions.

The dataset consisted of three sets of web documents to achieve an exhaustive evaluation of the system. The first set consisted of 152 website privacy policies. We collected 80 of them from Amazon Alexa's top 100 websites list for 2016 and 72 from the top two websites of each top Google Trends entry of 2017. Privacy policies of various companies have different HTML structure. They tend to contain many sections with each section having a title and corresponding prose text. They are also lengthy and include essential information which is often neglected by most users (Wilson et al., 2016). These factors make privacy policies ideal candidates for testing ASDUS. For the second set, we used www.randomlists.com to generate a list of 200 random sites. We selected the websites which had terms of service in English, and this left us with 100 documents. Similar to privacy policies, terms of service documents offer a reliable set of testing opportunities. For the third set, we wanted greater diversity in content and domain, leading us to chose web pages by collecting the top two to four Google search results for the following topics: *news, sports, botany, web design, photography, data science, cookie policies, HTML, history, migraine, dataset, technical documentation, shoes, grammar, kids stories and cricket*. We skipped web pages which did not have sectional demarcations. Table 1 has more details on the size of the dataset. The dataset for the DD approach consisted of the same 152 web privacy policies which were used in the DI approach. Privacy policies of different websites have content related to similar topics which makes them specific and suitable for the DD approach. Out of 152, 122 were used for training and development and the remaining 30 were used for testing. We manually annotated the entire data set and created sanitized versions out of them for evaluation.

We evaluated our models in two ways. One was the ability to detect all title and prose segments, measured via precision, recall and F-1 scores. This metric is evaluated by comparing the simplified output of ASDUS with the sanitized version. The output was deemed correct only when the system detected and produced both the title and its corresponding prose section. The second facet of evaluation was to determine the percentage of legitimate original text reproduced in the output. We name this the *coverage* of the output. A higher coverage indicates lower loss of text, which is desirable.

The results of the DI approach are presented in Table 2. The near-perfect recall is due to the ro-

Type	P	R	F1	S	C
PP	0.75	0.97	0.85	3611	96.92%
TOS	0.92	0.99	0.95	2299	98.85%
Misc	0.89	0.98	0.93	685	95.86%
Avg	0.82	0.98	0.89	6595	97.21%

Table 2: Results of the DI model. Columns represent type of web document, precision, recall, F-1 score, support and coverage respectively. Under *Type*, PP=privacy policies, TOS=terms of service, Misc=miscellaneous and Avg=weighted average.

bust method of learning XPaths of all titles coupled with the clustering method, which ensured the detection of nearly all the segments. Precision is slightly lower because some prose texts with relatively short lengths were wrongly classified as titles. The feature set enables the creation of two distinct clusters, which in turn results in all titles being detected during the clustering phase. To test the effect of each feature on the final results, an ablation study was conducted by removing one different feature from each run. This resulted in a total of seven runs, whose results are depicted in Table 3. The dropped feature is listed in the first column. The biggest drop in performance occurs when the own text length feature is dropped. The drop in F-1 scores in all runs suggests the contribution of every feature towards the result. The removal of discourse markers and named entity slots resulted in the least decrease in performance, and the greatest decrease came from removing text length.

Dropped Feature	P	R	F1
text length	0.69	0.73	0.71
stop words	0.75	0.75	0.75
punctuation symbols	0.76	0.79	0.77
sentences	0.76	0.79	0.77
next text length	0.76	0.81	0.78
capitalized first letter	0.78	0.80	0.79
discourse markers	0.79	0.85	0.81
named entity slots	0.79	0.85	0.81
None	**0.82**	**0.98**	**0.85**

Table 3: Results of the ablation study for the DI approach. The columns indicate dropped feature, precision, recall and F-1 score respectively.

In sharp contrast to the DI approach, the DD approach (Table 4) has high precision owing to the similarity of title texts across documents in the same domain and differences between vocabularies of title and prose text. Training on word embeddings of titles has rendered the system sensitive to variations, resulting in the rejection of many legitimate titles, which in turn led to the slightly lower recall and coverage values. This over-fitting can be mitigated by training on a larger corpus and by increasing the context window while generating word embeddings of prose text.

Precision	Recall	F-1	Supp.	Cover
0.99	0.75	0.86	754	93.10%

Table 4: Results of the DD model.

The complementary nature of the two models is identifiable from the results. The word embeddings of the DD approach contribute towards precision, and the lexical and morphological features of the DI approach contribute towards recall. We can treat the word embeddings as the semantic aspect of the underlying text and the lexical and morphological characteristics of the DI model as the syntactic aspect. Both of these aspects of language thus inform the orthographic structure of documents. However, the domain-independence or the ability of the DI approach to work on different domains without the need for prior labeled data is a substantial advantage over the DD approach, which requires pre-labeled data. Annotation of HTML documents and generation of the sanitized version is a tedious process and requires human effort, lending value to both methods.

5 Future Work and Conclusion

The two methods presented in this paper were effective in identifying the title and prose texts of segments of HTML pages. The DI approach, with its high coverage, is desirable when the penalty for losing text is high. The DD approach can be used in situations where precision is prioritized over recall. In the future, we intend to build upon our methods and enable automatic detection of subheaders. This would lead to the identification of hierarchical organization of the text, which provides a novel approach to generate web ontologies. Further, we have planned to generate titles for prose text essentially creating micro-summaries of text, a relatively unexplored area.

Acknowledgment

The work reported herein was conducted as part of the Usable Privacy Policy Project (Sadeh et al., 2013) and funded by the National Science Foundation under grant CNS-1330596.

References

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al. 2015. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*.

Nilesh Dalvi, Ravi Kumar, and Mohamed Soliman. 2011. Automatic wrappers for large scale web extraction. *Proceedings of the VLDB Endowment*, 4(4):219–230.

Tanya Denver. 2018. Discourse Markers – Connectors a list of discourse markers with examples. https://www.academia.edu/6888756/discourse_markers_connectors_a_list_of_discourse_markers_with_examples.

Sergio Flesca, Giuseppe Manco, Elio Masciari, Eugenio Rende, and Andrea Tagarelli. 2004. Web wrapper induction: A brief survey. *AI Communications*, 17(2):57–61.

Dayne Freitag. 1998. Information extraction from HTML: Application of a general machine learning approach. In *AAAI/IAAI*, pages 517–523.

Alberto P García-Plaza, Víctor Fresno, Raquel Martínez Unanue, and Arkaitz Zubiaga. 2017. Using fuzzy logic to leverage HTML markup for web page representation. *IEEE Transactions on Fuzzy Systems*, 25(4):919–933.

Yufan Guo, Roi Reichart, and Anna Korhonen. 2013. Improved information structure analysis of scientific documents through discourse and lexical constraints. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 928–937.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. 2009. The weka data mining software: An update. *ACM SIGKDD explorations newsletter*, 11(1):10–18.

Jonathan Hedley. 2017. jsoup (1.11.3). https://jsoup.org/.

Kenji Hirohata, Naoaki Okazaki, Sophia Ananiadou, and Mitsuru Ishizuka. 2008. Identifying sections in scientific abstracts using conditional random fields. In *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-I*.

Chenyi Liao, Kei Hiroi, Katsuhiko Kaji, and Nobuo Kawaguchi. 2015. An event data extraction method based on HTML structure analysis and machine learning. In *Computer Software and Applications Conference (COMPSAC), 2015 IEEE 39th Annual*, volume 3, pages 217–222. IEEE.

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA. http://is.muni.cz/publication/884893/en.

Norman Sadeh, Alessandro Acquisti, Travis D. Breaux, Lorrie Faith Cranor, Aleecia M. McDonald, Joel R. Reidenberg, Noah A. Smith, Fei Liu, N Cameron Russell, Florian Schaub, et al. 2013. The Usable Privacy Policy Project. Technical report, Technical report, Technical Report, CMU-ISR-13-119, Carnegie Mellon University.

Simone Teufel and Min-Yen Kan. 2011. Robust argumentative zoning for sensemaking in scholarly documents. In *Advanced Language Technologies for Digital Libraries*, pages 154–170. Springer.

Simone Teufel et al. 1999. *Argumentative zoning: Information extraction from scientific text*. Ph.D. thesis, University of Edinburgh.

Shomir Wilson. 2013. Toward automatic processing of english metalanguage. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 760–766.

Shomir Wilson and Jon Oberlander. 2014. Determiner-established deixis to communicative artifacts in pedagogical text. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 409–414.

Shomir Wilson, Florian Schaub, Aswarth Abhilash Dara, Frederick Liu, Sushain Cherivirala, Pedro Giovanni Leon, Mads Schaarup Andersen, Sebastian Zimmeck, Kanthashree Mysore Sathyendra, N. Cameron Russell, et al. 2016. The creation and analysis of a website privacy policy corpus. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1330–1340.