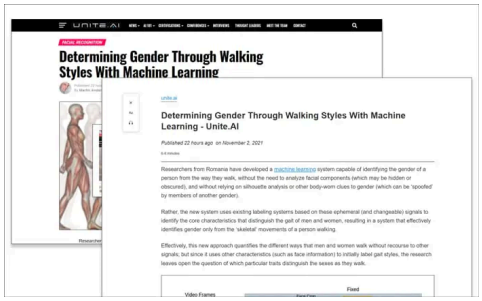
Beyond ‘Reader Mode’ With Machine Learning

Originally published at https://www.unite.ai/beyond-reader-mode-with-machine-learning/

Researchers from South Korea have used machine learning to develop an improved method for extracting actual content from web pages so that the ‘furniture’ of a web page – such as sidebars, footers and navigation headers, as well as advertisement blocks – disappears for the reader.

Though such functionality is either built into most popular web browsers, or else is easily available via extensions and plugins, these technologies rely on semantic formatting that may not be present in the web page, or which may have been deliberately compromised by the site owner in order to prevent the reader hiding the ‘full fat’ experience of the page.



One of our own web pages ‘slimmed down’ with Firefox’s integral Reader View functionality.

Instead, the new method uses a grid-based system that iterates through the web page, evaluating how pertinent the content is to the core aim of the page.

The content extraction pipeline first divides the page into a grid (upper row) before evaluating the relationship of found pertinent cells to other cells (middle) and finally merging the approved cells (bottom). Source: https://arxiv.org/ftp/arxiv/papers/2110/2110.14164.pdf

Once a pertinent cell is identified, its relationship with nearby cells is also evaluated before being merged into the interpreted ‘core content’.

The central idea of the approach is to abandon code-based markup as an index of relevance (i.e. HTML tags that would normally denote the beginning of a paragraph, for instance, which can be replaced by alternate tags that will 'fool' screen readers and utilities such as Reader View), and deduce the content based solely on its visual appearance.

The approach, called Grid-Center-Expand (GCE), has been extended by the researchers into Deep Neural Network (DNN) models that exploit Google's TabNet, an interpretative tabular learning architecture.

Get To the Point

The paper is titled *Don't read, just look: Main content extraction from web pages using visually apparent features*, and comes from three researchers at Hanyang University, and one from the Institute of Convergence Technology, all located in Seoul.

Improved extraction of core web page content is potentially valuable not only for the casual end-user, but also for machine systems that are tasked with ingesting or indexing domain content for the purposes of Natural Language Processing (NLP), and other sectors in AI.

As it stands, if non-relevant content is included in such extraction processes, it may need to be manually filtered (or labeled), at great expense; worse, if the unwanted content is included with the core content, it could affect how the core content is interpreted, and the outcome of transformer and encoder/decoder systems that are relying on clean content.

An improved method, the researchers argue, is especially necessary because existing approaches often fail with non-English web pages.

French, Japanese and Russian web pages are noted as scoring worst in success rates for the four most common 'Reader View' approaches: Mozilla's Readability.js; Google's DOM Distiller; Web2Text; and Boilernet.

Datasets and Training

The researchers compiled dataset material from English keywords in the GoogleTrends-2017 and GoogleTrends-2020 dataset, though they observe that, in terms of results, there were no practical differences between the two datasets.

Additionally, the authors gathered non-English keywords from South Korea, France, Japan, Russia, Indonesia and Saudi Arabia. Chinese keywords were added from a Baidu dataset, since Google Trends could not offer Chinese data.

Testing and Results

In testing the system, the authors found that it offer the same level of performance as recent DNN models, while providing better accommodation for a wider variety of languages.

For instance, the Boilernet architecture, while maintaining good performance in extracting pertinent content, adapts poorly to Chinese and Japanese datasets, while Web2Text, the authors find, has 'relatively poor performance' all round, with linguistic features that are not multilingual, and are unsuited for extracting central content from web pages.

Mozilla's Readability.js was found to achieve acceptable performance across multiple languages including English, even as a rule-based method. However the researchers found that its performance dropped notably on Japanese and French datasets, highlighting the limitations of trying to parse characteristics of a specific region entirely by rule-based approaches.

Meanwhile Google's DOM Distiller, which blends heuristics and machine learning approaches, was found to perform well across the board.

Table 3: LCS $F_{0.5}$ Score

	English (Global)		Non-English (Local, 2020 year only)							Average		
	2017	2020	KR	JP	CN	FR	SA	ID	RU	EN	Non-EN	TOTAL
GCE	**0.744**	0.762	0.842	0.706	**0.910**	**0.840**	0.714	0.802	0.795	**0.753**	**0.801**	**0.791**
Readability.js	0.728	**0.765**	0.755	0.609	0.759	0.697	**0.764**	0.814	0.766	0.746	0.738	0.740
DOM Distiller	0.722	0.729	0.741	**0.713**	0.857	0.742	0.722	0.831	0.805	0.726	0.773	0.762
Boilernet	0.706	0.706	**0.861**	0.452	0.654	0.712	0.663	**0.842**	**0.796**	0.706	0.711	0.710
Web2Text	0.591	0.620	0.706	0.679	0.881	0.676	0.646	0.709	0.676	0.605	0.710	0.687
Tabnet_basic	0.596	0.624	0.632	0.511	0.776	0.666	0.673	0.750	0.714	0.610	0.675	0.660
Tabnet_both	0.640	0.665	0.751	0.531	0.789	0.686	0.677	0.798	0.746	0.653	0.711	0.698

Table 4: Text Block matching $F_{0.5}$ Score

	English (Global)		Non-English (Local, 2020 year only)							Average		
	2017	2020	KR	JP	CN	FR	SA	ID	RU	EN	Non-EN	TOTAL
GCE	0.658	**0.692**	0.751	**0.646**	**0.842**	**0.781**	0.616	0.703	**0.737**	0.675	**0.725**	**0.714**
Readability.js	0.570	0.626	0.513	0.435	0.520	0.515	0.632	0.604	0.691	0.598	0.559	0.567
DOM Distiller	0.436	0.494	0.572	0.438	0.723	0.400	0.480	0.426	0.521	0.465	0.509	0.499
Boilernet	**0.693**	0.676	**0.867**	0.406	0.548	0.711	0.600	**0.741**	0.731	**0.685**	0.658	0.664
Web2Text	0.433	0.620	0.785	0.541	0.804	0.673	**0.642**	0.650	0.681	0.527	0.682	0.648
Tabnet_basic	0.560	0.585	0.621	0.395	0.659	0.587	0.595	0.691	0.651	0.572	0.600	0.594
Tabnet_both	0.603	0.624	0.693	0.421	0.669	0.595	0.621	0.723	0.686	0.613	0.630	0.626

Table of results for methods tested during the project, including the researchers' own GCE module. Higher numbers are better.

The researchers conclude that *'GCE does not need to keep up with the rapidly changing web environment because it relies on human nature—genuinely global and multilingual features'*.

Writer on machine learning, domain specialist in human image synthesis. Former head of research content at Metaphysic.ai, until its dissolution into DNEG's Brahma.ai.
Portfolio site: martinanderson.ai
Contact: martin@martinanderson.ai