An Empirical Comparison of Web Content Extraction Algorithms

Janek Bevendorff  Sanket Gupta  Johannes Kiesel  Benno Stein

Leipzig University, Bauhaus-Universität Weimar
Intelligent Information Systems Group – webis.de

SIGIR 2023 – July 23–27, Taipei, Taiwan
Web (Main) Content Extraction

What?

- Extraction of textual content from web pages, excluding navigation, ads, banners, etc.
- Also referred to as “boilerplate removal.”
- Extract is unstructured, (ideally coherent) running text.
Web (Main) Content Extraction

What?
- Extraction of textual content from web pages, excluding navigation, ads, banners, etc.
- Also referred to as “boilerplate removal.”
- Extract is unstructured, (ideally coherent) running text.

Why?
- Indexing / snippet generation for search applications.
- Content summarization.
- Assistive technologies.
- Training of (large) language models.
Europe heatwave: More record temperatures expected

By Robert Greenall
BBC news

Much of southern Europe is baking in extreme heat, with Greece seeing temperatures of 40°C (104°F) or more.

The Acropolis, the country's most popular tourist attraction, was closed during the hottest hours of the day to protect visitors.

Potentially record temperatures are expected next week as another heatwave approaches.

The European Space Agency (ESA) says Italy, Spain, France, Germany and Poland may see extreme conditions.

The ESA monitors land and sea temperatures via its satellites.

The hottest temperature ever recorded in Europe was 48.8°C in Sicily in August 2021.
Europe heatwave: More record temperatures expected

By Robert Greenall
BBC News

Much of southern Europe is baking in extreme heat, with Greece seeing temperatures of 48°C (118°F) or more.

The Acropolis, the country’s most popular tourist attraction, was closed during the hottest hours of the day to protect visitors.

Potentially record temperatures are expected next week as another heatwave approaches.

The European Space Agency (ESA) says Italy, Spain, France, Germany and Poland may see extreme conditions.

The ESA monitors land and sea temperatures via its satellites.

The hottest temperature ever recorded in Europe was 48°C in Sicily in August 2021.
2 Language Standards Supported by GCC

For each language compiled by GCC for which there is a standard, GCC attempts to follow one or more versions of that standard, possibly with some exceptions, and possibly with some extensions.

- C Language
- C++ Language
- Objective-C and Objective-C++ Languages
- Go Language
- D language
- References for Other Languages

2.1 C Language

The original ANSI C standard (X3.159-1989) was ratified in 1989 and published in 1990. This standard was ratified as an ISO standard (ISO/IEC 9899:1990) later in 1990. There were no technical differences between these publications, although the sections of the ANSI standard were renumbered and became clauses in the ISO standard. The ANSI standard, but not the ISO standard, also came with a Rationale document. This standard, in both its forms, is commonly known as C99, or occasionally as C00, from the dates of ratification. To select this standard in GCC, use one of the options -ansi-std-c99 or -std=iso9899:1990; to obtain all the diagnostics required by the standard, you should also specify -pedantic (or -pedantic-errors if you want them to be errors rather than warnings). See Options Controlling C Dialect.

Errors in the 1990 ISO C standard were corrected in two Technical Corrigenda published in 1994 and 1996. GCC does not support the uncorrected version.

An amendment to the 1990 standard was published in 1995. This amendment added digraphs and __STDC_VERSION__ to the language, otherwise concerned the library. This amendment is commonly known as AMDI; the amended standard is sometimes known as C94 or C95. To select this standard in GCC, use the option -std=iso9899:199449 (with, as for other standard versions, -pedantic to receive all required diagnostics).

A new edition of the ISO C standard was published in 1999 as ISO/IEC 9899:1999, and is commonly known as C99. (While in development, drafts of this standard version were referred to as C2X.) GCC has substantially complete support for this standard version; see https://gcc.gnu.org/c99status.html for details. To select this standard, use -std=c99 or -std=iso9899:1999.

Errors in the 1999 ISO C standard were corrected in three Technical Corrigenda published in 2001, 2004 and 2007. GCC does not support the uncorrected version.

A fourth version of the C standard, known as C11, was published in 2011 as ISO/IEC 9899:2011. (While in development, drafts of this standard version were referred to as C1X.) GCC has substantially complete support for this standard, enabled with -std=c11 or -std=iso9899:2011. A version with corrections integrated was prepared in 2017 and published in 2018 as ISO/IEC 9899:2018; it is known as C17 and is supported with -std=c17 or -std=iso9899:2017; the corrections are also applied with -std=c11, and the only difference between the options is the value of __STDC_VERSION__.

A further version of the C standard, known as C2X, is under development; experimental and incomplete support for this is enabled with -std=c2x.
2 Language Standards Supported by GCC

For each language compiled by GCC for which there is a standard, GCC attempts to follow one or more versions of that standard, possibly with some exceptions, and possibly with some extensions.

- C Language
- C++ Language
- Objective-C and Objective-C++ Languages
- Go Language
- D language
- References for Other Languages

2.1 C Language

The original ANSI C standard (X3.159-1989) was ratified in 1989 and published in 1990. This standard was ratified as an ISO standard (ISO/IEC 9899:1990) later in 1990. There were no technical differences between these publications, although the sections of the ANSI standard were renumbered and became clauses in the ISO standard. The ANSI standard, but not the ISO standard, also came with a Rationale document. This standard, in both its forms, is commonly known as C89, or occasionally as C90, from the dates of ratification. To select this standard in GCC, use one of the options -ansi, -std=c99 or -std=iso9899:1990, to obtain all the diagnostics required by the standard, you should also specify -pedantic (or -pedantic-errors if you want them to be errors rather than warnings). See Options Controlling C Dialect.

Errors in the 1990 ISO C standard were corrected in two Technical Corrigenda published in 1994 and 1996. GCC does not support the uncorrected version.

An amendment to the 1990 standard was published in 1995. This amendment added digraphs and __STDC_VERSION__ to the language, otherwise concerned the library. This amendment is commonly known as AM01; the amended standard is sometimes known as C94 or C95. To select this standard in GCC, use the option -std=iso9899:1994<95 (with, as for other standard versions, -pedantic to receive all required diagnostics).

A new edition of the ISO C standard was published in 1999 as ISO/IEC 9899:1999, and is commonly known as C99. (While in development, drafts of this standard version were referred to as C9X.) GCC has substantially complete support for this standard version; see https://gcc.gnu.org/c99status.html for details. To select this standard, use -std=c99 or -std=iso9899:1999.

Errors in the 1999 ISO C standard were corrected in three Technical Corrigenda published in 2001, 2004 and 2007. GCC does not support the uncorrected version.

A fourth version of the C standard, known as C11, was published in 2011 as ISO/IEC 9899:2011. (While in development, drafts of this standard version were referred to as C1X.) GCC has substantially complete support for this standard, enabled with -std=c11 or -std=iso9899:2011. A version with corrections integrated was prepared in 2017 and published in 2018 as ISO/IEC 9899:2018; it is known as C17 and is supported with -std=c17 or -std=iso9899:2017; the corrections are also applied with -std=c11, and the only difference between the options is the value of __STDC_VERSION__.

A further version of the C standard, known as C2X, is under development; experimental and incomplete support for this is enabled with -std=c2x.
Goldstandard datasets are rare (and also small and quite dated).

Academic research and open source tools exist.

Yet, very little rigorous and comparable evaluation has been done.

There is no clear-cut definition of “main content.”
Web Content Extraction

Datasets

We collected, cleaned, and combined 8 datasets of varying complexity:

- CETD (700)
- CleanEval (738)
- CleanPortalEval (71)
- Dragnet (1,379)
- Google-Trends (180)
- L3S-GN1 (621)
- Readability (115)
- Scrapinghub (181)
- (Combined: 3,985 pages)
Web Content Extraction

Datasets

We collected, cleaned, and combined 8 datasets of varying complexity:

- CETD (700)
- CleanEval (738)
- CleanPortalEval (71)
- Dragnet (1,379)
- Google-Trends (180)
- L3S-GN1 (621)
- Readability (115)
- Scrapinghub (181)
- (Combined: 3,985 pages)

\[
c = 1 - \frac{|\{ t \in T : \text{truth}(t) = 1 \}|}{|T|}
\]
Web Content Extraction

Datasets

We collected, cleaned, and combined 8 datasets of varying complexity:

- CETD (700)
- CleanEval (738)
- CleanPortalEval (71)
- Dragnet (1,379)
- Google-Trends (180)
- L3S-GN1 (621)
- Readability (115)
- Scrapinghub (181)
- (Combined: 3,985 pages)
We reproduced **14 main content extractors** and **5 baseline text converters**:

<table>
<thead>
<tr>
<th>Heuristic</th>
<th>ML-based</th>
<th>Baseline</th>
</tr>
</thead>
<tbody>
<tr>
<td>BTE</td>
<td>BoilerNet</td>
<td>BS4</td>
</tr>
<tr>
<td>Goose3</td>
<td>Boilerpipe</td>
<td>html_text</td>
</tr>
<tr>
<td>jusText</td>
<td>Dragnet</td>
<td>inscriptis</td>
</tr>
<tr>
<td>Newspaper3k</td>
<td>ExtractNet</td>
<td>lxml Cleaner</td>
</tr>
<tr>
<td>Readability</td>
<td>Go DOM Distiller</td>
<td>XPath Text</td>
</tr>
<tr>
<td>Resiliparse</td>
<td>news-please</td>
<td></td>
</tr>
<tr>
<td>Trafilatura</td>
<td>Web2Text</td>
<td></td>
</tr>
</tbody>
</table>
Web Content Extraction
SotA Extraction Systems

We reproduced 14 main content extractors and 5 baseline text converters:

<table>
<thead>
<tr>
<th>Heuristic</th>
<th>ML-based</th>
<th>Baseline</th>
</tr>
</thead>
<tbody>
<tr>
<td>BTE</td>
<td>BoilerNet</td>
<td>BS4</td>
</tr>
<tr>
<td>Goose3</td>
<td>Boilerpipe</td>
<td>html_text</td>
</tr>
<tr>
<td>jusText</td>
<td>Dragnet</td>
<td>inscriptis</td>
</tr>
<tr>
<td>Newspaper3k</td>
<td>ExtractNet</td>
<td>lxml Cleaner</td>
</tr>
<tr>
<td>Readability</td>
<td>Go DOM Distiller</td>
<td>XPath Text</td>
</tr>
<tr>
<td>Resiliparse</td>
<td>news-please</td>
<td></td>
</tr>
<tr>
<td>Trafilatura</td>
<td>Web2Text</td>
<td></td>
</tr>
</tbody>
</table>

Three 4-gram majority vote ensembles (66%): All / Best only / Best only (weighted)
Extractor performance ranked by ROUGE-L\textsuperscript{1} at summary level (ROUGE-LSum):

\textit{LCS} : (Union) Longest Common Subsequence

The quick \textit{brown fox} jumps over the \textit{lazy dog}. \quad A \textit{brown fox} hits the \textit{crazy dog}.

\textsuperscript{1}Lin, 2004; ROUGE: A Package for Automatic Evaluation of Summaries
Extractor performance ranked by ROUGE-L\(^1\) at summary level (ROUGE-LSum):

\(LCS\) : (Union) Longest Common Subsequence

The quick \textbf{brown fox} jumps over the \textbf{lazy dog}. A \textbf{brown fox} hits the \textbf{crazy dog}.

\[
P_{lcs} = \frac{\sum_i^n LCS(T_i, C)}{|C|_{\text{words}}}, \quad R_{lcs} = \frac{\sum_i^n LCS(T_i, C)}{|T|_{\text{words}}}.
\]

\(T\) : Target sentence(s)
\(C\) : Candidate sentence(s)

\(^1\)Lin, 2004; ROUGE: A Package for Automatic Evaluation of Summaries
Web Content Extraction
Evaluation – All Pages

ROUGE-LSum Median $F_1$ Page Scores

(Best weighted) (Best only) ( Majority all) Readability Trafilatura DOM Distiller news-please Newspaperck Boilerpipe Dragnet Goose3 Resiliparse BTE Web2Text JustText ExtractNet BoilerNet xml Cleaner html_text BS4 inscriptis XPath Text

0.00
0.25
0.50
0.75
1.00

16 July 2023 • webis.de
Web Content Extraction

Evaluation – Complex Pages ($Q_4 < c \leq 1$)

ROUGE-LSum Median $F_1$ Page Scores
Web Content Extraction
Evaluation – Easy Pages ($0 \leq c < Q_1$)

ROUGE-LSum Median $F_1$ Page Scores

(Best weighted) (Best only) (Majority all)

<table>
<thead>
<tr>
<th>Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Web2Text</td>
</tr>
<tr>
<td>Readability</td>
</tr>
<tr>
<td>DOM Distiller</td>
</tr>
<tr>
<td>Resiliparse</td>
</tr>
<tr>
<td>BTE</td>
</tr>
<tr>
<td>Trafiliatura</td>
</tr>
<tr>
<td>jusText</td>
</tr>
<tr>
<td>html_text</td>
</tr>
<tr>
<td>lxml Cleaner</td>
</tr>
<tr>
<td>inscriptis</td>
</tr>
<tr>
<td>Newspaper3k</td>
</tr>
<tr>
<td>news-please</td>
</tr>
<tr>
<td>ExtractNet</td>
</tr>
<tr>
<td>BoilerNet</td>
</tr>
<tr>
<td>BS4</td>
</tr>
<tr>
<td>Dragnet</td>
</tr>
<tr>
<td>Boilerpipe</td>
</tr>
<tr>
<td>XPath Text</td>
</tr>
<tr>
<td>Goose3</td>
</tr>
</tbody>
</table>

0.00 0.25 0.50 0.75 1.00

July 2023 • webis.de
Summary

- New datasets needed!
- Precision-oriented heuristic models perform quite well.
- Deep neural models perform rather poorly (so far).
- Please don't report only single numbers!
- Readability / Trafilatura / DOM Distiller are the most robust right now.
- Resiliparse (ours) is (not yet!) the best, but the fastest by an order of magnitude. :-(
Summary

- New datasets needed!
- Precision-oriented heuristic models perform quite well.
- Deep neural models perform rather poorly (so far).
- Please don't report only single numbers!
- Readability / Trafilatura / DOM Distiller are the most robust right now.
- Resiliparse (ours) is (not yet!) the best, but the fastest by an order of magnitude. :-)

More in our paper, all code and data publicly available:

- github.com/webis-de/SIGIR-23
- webis.de/publications