Boilerplate Detection using Shallow Text Features

Christian Kohlschütter, Peter Fankhauser, Wolfgang Nejdl
The L3S Research Center focuses on fundamental and application-oriented research in all areas of Web Science. L3S researchers develop new methods and technologies that enable intelligent, seamless access to information via the Web; link individuals and communities in all areas of the knowledge society, including academia and education; and connect the Internet to the real world.

In the context of a large number of projects, the L3S explores numerous issues covering the entire spectrum of challenges in Web Science as a field of research. Since its founding in 2001, the L3S has brought together numerous scholars and researchers who actively take on these challenges and perform interdisciplinary research in the fields of information retrieval, databases, the Semantic Web, performance modeling, service computing, and mobile networks. The center's total research volume is more than 6 million euros per year, with a large number of projects in the areas of

- Intelligent Access to Information
- Next Generation Internet
- e-Science

The L3S is a research-driven institution that attracts outstanding students and researchers from all over the world with its open and invigorating research culture. For young researchers, the L3S is encouraging, innovative, international, independent, and supportive.

L3S activities primarily focus on research, but also include consulting and technology transfer. This is made possible by complementary background knowledge that L3S researchers themselves bring to their work, and the center's collaborations and projects with scholars and researchers not only from computer sciences, but also including library sciences, linguistics, psychology, law, economics, and business administration.

The experience L3S has gained over the years in participating in a variety of projects funded by the European Union has led to a large number of cooperations with research institutions and companies throughout all of Europe, and in many research results and products. Since 2008 alone, the L3S has been involved in 12 EU projects as part of the EU's Seventh Framework Programme, four of them (LivingKNOWedge, Okkam, EUWB and EERQI) integrated projects, as well as the STELLAR Network of Excellence.

In addition to its international cooperations, with its interdisciplinary research initiative entitled "Future Internet - Internet, Information and I," L3S is playing a key role in the development of this important topic for the future of Lower Saxony as well.
The Advisory Board visiting L3S Research Center

L3S Research Center

The L3S Research Center focuses on fundamental and application-oriented research in all areas of Web Science. L3S researchers develop new methods and technologies that enable intelligent, seamless access to information via the Web; link individuals and communities in all areas of the knowledge society, including academia and education; and connect the Internet to the real world.

In the context of a large number of projects, the L3S explores numerous issues covering the entire spectrum of challenges in Web Science as a field of research. Since its founding in 2001, the L3S has brought together numerous scholars and researchers who actively take on these challenges and perform interdisciplinary research in the fields of information retrieval, databases, the Semantic Web, performance modeling, service computing, and mobile networks. The center’s total research volume is more than 6 million euros per year, with a large number of projects in the areas of

- Intelligent Access to Information
- Next Generation Internet
- E-Science

The L3S is a research-driven institution that attracts outstanding students and researchers from all over the world with its open and invigorating research culture. For young researchers, the L3S is encouraging, innovative, international, independent, and supportive.

L3S activities primarily focus on research, but also include consulting and technology transfer. This is made possible by complementary background knowledge that L3S researchers themselves bring to their work, and the center’s cooperation and projects with scholars and researchers not only from computer sciences, but also including library sciences, linguistics, psychology, law, economics, and business administration.

The experience L3S has gained over the years in participating in a variety of projects financed by the European Union has led to a large number of cooperations with research institutions and companies throughout all of Europe, and in many research results and products. Since 2008 alone, the L3S has been involved in 12 EU projects as part of the EU’s Seventh Framework Programme, four of them (LivingKnowledge, Oikos, EUWB and EDIGI) integrated projects, as well as the STELLAR Network of Excellence.

In addition to its international cooperations, with its interdisciplinary research initiative entitled "Future Internet - Internet, Information and I," L3S is playing a key role in the development of this important topic for the future of Lower Saxony as well.
The L3S Research Center focuses on fundamental and application-oriented research in all areas of Web Science. L3S researchers develop new methods and technologies that enable intelligent, automated access to information via the Web, link individuals and communities in all areas of the knowledge society, including academia and education; and connect the Internet to the real world.

In the context of a large number of projects, the L3S explores numerous issues covering the entire spectrum of challenges in Web Science as a field of research. Since its founding in 2001, the L3S has brought together numerous scholars and researchers who actively take on these challenges and perform interdisciplinary research in the fields of information retrieval, databases, the Semantic Web, performance modeling, service computing, and mobile networks. The center’s total research volume is more than 6 million euros per year, with a large number of projects in the areas of

- Intelligent Access to Information
- Next Generation Internet
- E-Science

The L3S is a research-driven institution that attracts outstanding students and researchers from all over the world with its open and invigorating research culture. For young researchers, the L3S is encouraging, innovative, international, independent, and supportive.

L3S activities primarily focus on research, but also include consulting and technology transfer. This is made possible by complementary background knowledge that L3S researchers themselves bring to their work, and the center’s cooperation and projects with scholars and researchers not only from computer sciences, but also including library sciences, linguistics, psychology, law, economics, and business administration.

The experience L3S has gained over the years in participating in a variety of projects financed by the European Union has led to a large number of cooperations with research institutions and companies throughout all of Europe, and in many research results and products. Since 2008 alone, the L3S has been involved in 12 EU projects as part of the EU’s Seventh Framework Programme, four of them (LivingKnowledge, Okkam, EUWB and EERQI) integrated projects, as well as the STElLaR Network of Excellence.

In addition to its international cooperations, with its interdisciplinary research initiative entitled “Future Internet – Internet, Information and I,” L3S is playing a key role in the development of this important topic for the future of Lower Saxony as well.
L3S Research Center

The L3S Research Center focuses on fundamental and application-oriented research in all areas of Web Science. L3S researchers develop new methods and technologies that enable intelligent, seamless access to information via the Web; link individuals and communities in all areas of the knowledge society, including academia and education; and connect the Internet to the real world.

In the context of a large number of projects, the L3S explores numerous issues covering the entire spectrum of challenges in Web Science as a field of research. Since its founding in 2001, the L3S has brought together numerous scholars and researchers who actively take on these challenges and perform interdisciplinary research in the fields of information retrieval, databases, the Semantic Web, performance modeling, service computing, and mobile networks. The center’s total research volume is more than 6 million euros per year, with a large number of projects in the areas of

* Intelligent Access to Information
* Next Generation Internet
* E-Science

In addition to its international cooperations, with its interdisciplinary research initiative entitled “Future Internet – Internet, Information and I,” L3S is playing a key role in the development of this important topic for the future of Lower Saxony as well.
The L3S Research Center focuses on fundamental and application-oriented research in all areas of Web Science. L3S researchers develop new methods and technologies that enable intelligent, seamless access to information via the Web; link individuals and communities in all areas of the knowledge society, including academia and education; and connect the Internet to the real world.

In the context of a large number of projects, the L3S explores numerous issues covering the entire spectrum of challenges in Web Science as a field of research. Since its founding in 2001, the L3S has brought together numerous scholars and researchers who actively take on these challenges and perform interdisciplinary research in the fields of information retrieval, databases, the Semantic Web, performance modeling, service computing, and mobile networks. The center’s total research volume is more than 6 million euros per year, with a large number of projects in the areas of:

- **Intelligent Access to Information**
- **Next Generation Internet**
- **E-Science**

The L3S is a research-driven institution that attracts outstanding students and researchers from all over the world with its open and invigorating research culture. For young researchers, the L3S is encouraging, innovative, international, independent, and supportive.

L3S activities primarily focus on research, but also include consulting and technology transfer. This is made possible by complementary background knowledge that L3S researchers themselves bring to their work, and the center’s cooperations and projects with scholars and researchers not only from computer sciences, but also including library sciences, linguistics, psychology, law, economics, and business administration.

The experience L3S has gained over the years in participating in a variety of projects financed by the European Union has led to a large number of cooperations with research institutions and companies throughout all of Europe, and in many research results and products. Since 2008 alone, the L3S has been involved in 12 EU projects as part of the EU’s Seventh Framework Programme, four of them (LivingKnowledge, Okkam, EUWB and EERQI) integrated projects, as well as the STELLAR Network of Excellence.

In addition to its international cooperations, with its interdisciplinary research initiative entitled “Future Internet – Internet, Information and I,” L3S is playing a key role in the development of this important topic for the future of Lower Saxony as well.
Existing Approaches

• Machine Learning vs. Heuristics
• Site-specific Solutions
  (Rule-based Scraping, DOM, Text, Link Graph)
• Vision-based models
• Tokens, N-Grams
• Shallow Text Features
• Context
Shallow Text Features

- Examine Document at Text Block Level
  - Numbers: Words, Tokens contained in block
  - Average Lengths: Tokens, Sentences
  - Ratios: Uppercased words, full stops
  - Classes: Block-level HTML tags <P>, <Hn>, <DIV>
  - Densities: Link Density (Anchor Text Percentage), Text Density
The L3S Research Center focuses on fundamental and application-oriented research in all areas of Web Science. L3S researchers develop new methods and technologies that enable intelligent, seamless access to information via the Web; link individuals and communities in all areas of the knowledge society, including academia and education; and connect the Internet to the real world.

In the context of a large number of projects, the L3S explores numerous issues covering the entire spectrum of challenges in Web Science in a field of research. Since its founding in 2001, the L3S has brought together numerous scholars and researchers who actively take on these challenges and perform interdisciplinary research in the fields of information retrieval, databases, the Semantic Web, performance modeling, service computing, and mobile networks. The center’s total research volume is more than 6 million euros per year, with a large number of projects in the areas of:

- Intelligent Access to Information
- Next Generation Internet
- E-Science

The L3S is a research-driven institution that attracts outstanding students and researchers from all over the world with its open and invigorating research culture. For young researchers, the L3S is encouraging, innovative, international, independent, and supportive.

L3S activities primarily focus on research, but also include consulting and technology transfer. This is made possible by crossdisciplinary backgrounds. Besides, the L3S offers fascinating training in a new field that is characterized by teamwork, collaboration, and individual responsibility.

The experience L3S has gained over the years in participating in a variety of projects financed by the European Union has led to a large number of collaborations with research institutions and companies throughout the world, and has also resulted in a large number of publications in the areas of information retrieval, databases, and mobile networks.

In addition to its international cooperations, the L3S is actively working on the development of the future Internet, with the Future Internet - Information and Communication Technologies (ICT) Framework Programme (FP7) of the European Union.
Contextual Features

• Intra-Document:
  • Relative/Absolute Position of Block
  • Features of the previous/next block

• Inter-Document

• Text Block Frequency
Experiments

1. Classification Accuracy?
   Decision Trees, SVM, 10-fold cross validation, F-Measure/ROC AuC, ...

2. Main Content Extraction
   Compare to BTE (Finn et al., 2001) and n-grams (Pasternack et al., 2009)
   In Paper also: Victor (Spousta et al., 2008), NCleaner (Evert, 2008)

3. Ranking Improvement?
   Precision@10, NDCG@10
   50 top-k TREC-Queries for BLOGS06 (3M docs)
GoogleNews Dataset

- L3S-GN1

621 news articles from 408 web sites, randomly sampled from a 254,000 pages crawl of English Google News over 4 months, manually assessed by L3S colleagues.

<table>
<thead>
<tr>
<th>Class</th>
<th># Blocks</th>
<th># Words</th>
<th># Tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>72662</td>
<td>520483</td>
<td>644021</td>
</tr>
<tr>
<td>Boilerplate</td>
<td>79%</td>
<td>35%</td>
<td>46%</td>
</tr>
<tr>
<td>Any Content</td>
<td>21%</td>
<td>65%</td>
<td>54%</td>
</tr>
<tr>
<td>Headline</td>
<td>1%</td>
<td>1%</td>
<td>1%</td>
</tr>
<tr>
<td>Article Full-text</td>
<td>12%</td>
<td>51%</td>
<td>42%</td>
</tr>
<tr>
<td>Supplemental</td>
<td>3%</td>
<td>3%</td>
<td>2%</td>
</tr>
<tr>
<td>User Comments</td>
<td>1%</td>
<td>1%</td>
<td>1%</td>
</tr>
<tr>
<td>Related Content</td>
<td>4%</td>
<td>9%</td>
<td>8%</td>
</tr>
</tbody>
</table>
Classification Accuracy
Block-Level (weighted by number of words)

ZeroR (baseline; predict “Content”)
- F1: 0%
- ROC AuC: 0%

Only Avg. Sentence Length
- F1: 25%
- ROC AuC: 50%

C4.8 Element Frequency (P/C/N)
- F1: 50%
- ROC AuC: 75%

Only Avg. Word Length
- F1: 75%
- ROC AuC: 100%

Only Number of Words @15
- F1: 49.7%
- ROC AuC: 49%

Only Link Density @0.33
- F1: 68%
- ROC AuC: 68%

1R: Text Density @10.5
- F1: 73.3%
- ROC AuC: 73.8%

C4.8 Link Density (P/C/N)
- F1: 70.9%
- ROC AuC: 70.9%

C4.8 Number of Words (P/C/N)
- F1: 77.5%
- ROC AuC: 78.8%

C4.8 All Local Features (C)
- F1: 86.7%
- ROC AuC: 85.6%

C4.8 NumWords + LinkDensity, simplified
- F1: 87.4%
- ROC AuC: 84.3%

C4.8 Text + LinkDensity, simplified
- F1: 87.9%
- ROC AuC: 86.8%

C4.8 All Local Features (P/C/N)
- F1: 91%
- ROC AuC: 94.2%

C4.8 NumWords + LinkDensity, simplified
- F1: 94.7%
- ROC AuC: 94.7%

C4.8 Text + LinkDensity, simplified
- F1: 94.9%
- ROC AuC: 94.4%

C4.8 All Local Features (C) + TDQ
- F1: 97%
- ROC AuC: 97.2%

C4.8 Text+Link Density (P/C/N)
- F1: 97.6%
- ROC AuC: 97.6%

C4.8 All Link Features (P/C/N)
- F1: 95%
- ROC AuC: 98%

C4.8 All Local Features + Global Freq.
- F1: 96.6%
- ROC AuC: 96.6%

SMO All Local Features + Global Freq.
- F1: 98%
- ROC AuC: 98%

- NumLeaves
- NumFeatures
Classification Accuracy
Block-Level (weighted by number of words)

- ZeroR (baseline; predict “Content”)
- Only Avg. Sentence Length
- C4.8 Element Frequency (P/C/N)
- Only Avg. Word Length
- Only Number of Words @15
- Only Link Density @0.33
- 1R: Text Density @10.5
- C4.8 Link Density (P/C/N)
- C4.8 Number of Words (P/C/N)
- C4.8 All Local Features (C)
- C4.8 NumWords + Link Density, simplified
- C4.8 Text + LinkDensity, simplified
- C4.8 All Local Features (C) + TDQ
- C4.8 Text+Link Density (P/C/N)
- C4.8 All Local Features (P/C/N)
- C4.8 All Local Features + Global Freq.
- SMO All Local Features + Global Freq.

F-Measure
- NumWords + Link Density: 92.2%
- ROC AuC: 95.7%

Other models show different F-Measures and ROC AuCs, varying from 49.7% to 98.1% for F-Measure and 49% to 95% for ROC AuC.
Classification Accuracy
Block-Level (weighted by number of words)

- ZeroR (baseline; predict “Content”)
- Only Avg. Sentence Length
- Only Avg. Word Length
- Only Number of Words @15
- Only Link Density @0.33
- 1R: Text Density @10.5
- C4.8 Link Density (P/C/N)
- C4.8 Number of Words (P/C/N)
- C4.8 All Local Features (C)
- C4.8 NumWords + LinkDensity, simplified
- C4.8 Text + LinkDensity, simplified
- C4.8 All Local Features + Global Freq.
- SMO All Local Features + Global Freq.

- NumWords + Link Density: F-Measure 92.2% ROC AuC 95.7%
- Text Density + Link Density: F-Measure 92.4% ROC AuC 96.9%

F-Measure and ROC AuC values are shown for various feature sets and models.
Classification Accuracy
Block-Level (weighted by number of words)

- ZeroR (baseline; predict “Content”)
- Only Avg. Sentence Length
- C4.8 Element Frequency (P/C/N)
- Only Avg. Word Length
- Only Number of Words @15
- Only Link Density @0.33
- 1R: Text Density @10.5
- C4.8 Link Density (P/C/N)
- C4.8 Number of Words (P/C/N)
- C4.8 All Local Features (C)
- C4.8 NumWords + Link Density, simplified
- C4.8 Text + Link Density, simplified
- C4.8 All Local Features (C) + TDQ
- C4.8 Text + Link Density (P/C/N)
- C4.8 All Local Features (P/C/N)
- C4.8 All Local Features + Global Freq.
- SMO All Local Features + Global Freq.

NumWords + Link Density
- F-Measure: 92.2%
- ROC AuC: 95.7%

Text Density + Link Density
- F-Measure: 92.4%
- ROC AuC: 96.9%

All Local Features
- F-Measure: 95%
- ROC AuC: 98.1%
"Main Content" Extraction
"Main Content" Extraction

Token-Level F-Measure

\( \mu = 68.30\% ; m = 70.60\% \) Baseline (Keep everything)
"Main Content" Extraction

Token-Level F-Measure

- $\mu=78.65\%$; $m=87.19\%$ Pasternack Trigrams, trained on News Corpus
- $\mu=68.30\%$; $m=70.60\%$ Baseline (Keep everything)
"Main Content" Extraction

- $\mu=80.78\%; m=85.10\%$  Keep everything with $\geq 10$ words
- $\mu=78.65\%; m=87.19\%$  Pasternack Trigrams, trained on News Corpus
- $\mu=68.30\%; m=70.60\%$  Baseline (Keep everything)
"Main Content" Extraction

Token-Level F-Measure

- $\mu=89.29\%; m=96.28\%$ BTE
- $\mu=80.78\%; m=85.10\%$ Keep everything with $\geq 10$ words
- $\mu=78.65\%; m=87.19\%$ Pasternack Trigrams, trained on News Corpus
- $\mu=68.30\%; m=70.60\%$ Baseline (Keep everything)
"Main Content" Extraction

![Diagram showing Token-Level F-Measure across different methods and document counts]

- **Baseline (Keep everything)**: \( \mu = 68.30\% ; m = 70.60\% \)
- **Pasternack Trigrams, trained on News Corpus**: \( \mu = 78.65\% ; m = 87.19\% \)
- **Keep everything with >= 10 words**: \( \mu = 80.78\% ; m = 85.10\% \)
- **BTE**: \( \mu = 89.29\% ; m = 96.28\% \)
- **Densitometric Classifier**: \( \mu = 90.61\% ; m = 95.56\% \)
- **Pasternack Trigrams, trained on News Corpus**: \( \mu = 78.65\% ; m = 87.19\% \)
"Main Content" Extraction

![Graph showing Token-Level F-Measure vs. # Documents for various methods]

- **µ=91.08%; m=95.87%** NumWords/LinkDensity Classifier
- **µ=90.61%; m=95.56%** Densitometric Classifier
- **µ=89.29%; m=96.28%** BTE
- **µ=80.78%; m=85.10%** Keep everything with >= 10 words
- **µ=78.65%; m=87.19%** Pasternack Trigrams, trained on News Corpus
- **µ=68.30%; m=70.60%** Baseline (Keep everything)
"Main Content" Extraction

![Graph showing token-level F-Measure](image)

- Blue line with label: \(\mu=92.17\%; m=97.65\%\) NumWords/LinkDensity + Largest Content Filter
- Dashed line with label: \(\mu=92.08\%; m=97.62\%\) Densitometric Classifier + Largest Content Filter
- Solid line with label: \(\mu=91.08\%; m=95.87\%\) NumWords/LinkDensity Classifier
- Dotted line with label: \(\mu=90.61\%; m=95.56\%\) Densitometric Classifier
- Dashed-dotted line with label: \(\mu=89.29\%; m=96.28\%\) BTE
- Black line with label: \(\mu=80.78\%; m=85.10\%\) Keep everything with >= 10 words
- Dotted-dashed line with label: \(\mu=78.65\%; m=87.19\%\) Pasternack Trigrams, trained on News Corpus
- Solid line with label: \(\mu=68.30\%; m=70.60\%\) Baseline (Keep everything)
"Main Content" Extraction

[Graph showing token-level F-measure for different methods and datasets]

- \( \mu = 95.93\%; m = 98.66\% \) NumWords/LinkDensity + Main Content Filter
- \( \mu = 95.62\%; m = 98.49\% \) Densitometric Classifier + Main Content Filter
- \( \mu = 92.17\%; m = 97.65\% \) NumWords/LinkDensity + Largest Content Filter
- \( \mu = 92.08\%; m = 97.62\% \) Densitometric Classifier + Largest Content Filter
- \( \mu = 91.08\%; m = 95.87\% \) NumWords/LinkDensity Classifier
- \( \mu = 90.61\%; m = 95.56\% \) Densitometric Classifier
- \( \mu = 89.29\%; m = 96.28\% \) BTE
- \( \mu = 89.29\%; m = 96.28\% \) BTE
- \( \mu = 80.78\%; m = 85.10\% \) Keep everything with >= 10 words
- \( \mu = 78.65\%; m = 87.19\% \) Pasternack Trigrams, trained on News Corpus
- \( \mu = 68.30\%; m = 70.60\% \) Baseline (Keep everything)
NumWords + Link Density

curr_linkDensity $\leq 0.333333$
| prev_linkDensity $\leq 0.555556$
| | curr_numWords $\leq 16$
| | | next_numWords $\leq 15$
| | | | prev_numWords $\leq 4$: BOILERPLATE
| | | | prev_numWords $> 4$: CONTENT
| | | next_numWords $> 15$: CONTENT
| | curr_numWords $> 16$: CONTENT
| prev_linkDensity $> 0.555556$
| | curr_numWords $\leq 16$: CONTENT
| | curr_numWords $> 40$: CONTENT
| curr_linkDensity $> 0.333333$: BOILERPLATE

Text Density + Link Density

curr_linkDensity $\leq 0.333333$
| prev_linkDensity $\leq 0.555556$
| | curr_textDensity $\leq 9$
| | | next_textDensity $\leq 10$
| | | | prev_textDensity $\leq 4$: BOILERPLATE
| | | | prev_textDensity $> 4$: CONTENT
| | | next_textDensity $> 10$: CONTENT
| | curr_textDensity $> 9$
| | | next_textDensity $= 0$: BOILERPLATE
| | | next_textDensity $> 0$: CONTENT
| prev_linkDensity $> 0.555556$
| | next_textDensity $\leq 11$: BOILERPLATE
| | next_textDensity $> 11$: CONTENT
curr_linkDensity $> 0.333333$: BOILERPLATE
Bernoulli trial: Transition to next block is success $p$
emission of another word is failure $1 - p$

$$Pr(Y = k) = (1 - p)^k p$$

$$Pr(Y = x) = (1 - p)^{x-1} \cdot p = P_T(T)^{x-1} \cdot P_T(N)$$
Stratified Model

\[ Pr(Y = x) = P_N(S) \cdot [P_S(S)^{x-1} \cdot P_S(N)] + \]
\[ + P_N(L) \cdot [P_L(L)^{x-1} \cdot P_L(N)] \]

\( L = "Long \ Text" \)
\( S = "Short \ Text" \)
\( P_S(N) \gg P_L(N) \)
\( P_N(L) = 1 - P_N(S) \)
Stratified Model

\[ Pr(Y = x) = P_N(S) \cdot [P_S(S)^{x-1} \cdot P_S(N)] + \]
\[ + P_N(L) \cdot [P_L(L)^{x-1} \cdot P_L(N)] \]

\( R^2_{\text{adj}} = 98.8\% \)
\( \text{RMSE} = 0.0027 \)

\[ L = "\text{Long Text}" \]
\[ S = "\text{Short Text}" \]

\[ P_S(N) \gg P_L(N) \]
\[ P_N(L) = 1 - P_N(S) \]
Stratified Model

$1 + E = 1 + \frac{1}{p} = 23.8$

$P_L(N) = 0.0437$

$P_S(N) = 0.3968$

$1 + E = 1 + \frac{1}{p} = 3.52$

$R^2_{adj} = 98.8\%$

$RMSE = 0.0027$

$L = "Long Text"

$S = "Short Text"

$P_S(N) \gg P_L(N)$

$P_N(L) = 1 - P_N(S)$

$Pr(Y = x) = P_N(S) \cdot \left[ P_S(S)^{x-1} \cdot P_S(N) \right] +$

$+ P_N(L) \cdot \left[ P_L(L)^{x-1} \cdot P_L(N) \right]$
Stratified Model

1 + E = 1 + 1/p = 23.8
$P_L(N) = 0.0437$

$P_S(N) = 0.3968$
$1 + E = 1 + 1/p = 3.52$

$L = "Long Text"
$S = "Short Text"

$P_S(N) \gg P_L(N)$
$P_N(L) = 1 - P_N(S)$

$R^2_{adj} = 98.8\%$
$RMSE = 0.0027$

GoogleNews assessment:
79% of blocks were boilerplate

$Pr(Y = x) = P_N(S) \cdot \left[ P_S(S)^{x-1} \cdot P_S(N) \right] +$
$+ P_N(L) \cdot \left[ P_L(L)^{x-1} \cdot P_L(N) \right]$
Retrieval Experiment

Baseline: $P@10=0.18; NDCG@10=0.0985$
BTE: $P@10=0.33; NDCG@10=0.1627$

50 top-k TREC queries on BLOGS06 dataset (~3M docs)
Retrieval Experiment

Baseline: $P@10=0.18; \ NDCG@10=0.0985$

BTE: $P@10=0.33; \ NDCG@10=0.1627$

50 top-k TREC queries on BLOGS06 dataset (~3M docs)
Retrieval Experiment

Improvement over Baseline: 144%/151%
Improvement over BTE: 33%/52%

P@10=0.18; NDCG@10=0.0985
P@10=0.33; NDCG@10=0.1627

50 top-k TREC queries on BLOGS06 dataset (~3M docs)
Conclusions
Conclusions

- Text Creation can be modeled as a Stratified Stochastic Process
Conclusions

• Text Creation can be modeled as a Stratified Stochastic Process

• Very high Classification/Extraction Accuracy (92-98%) at almost no cost
Conclusions

• Text Creation can be modeled as a Stratified Stochastic Process

• Very high Classification/Extraction Accuracy (92-98%) at almost no cost

• Increase of Retrieval Precision (33%-151%) at almost no cost
Next Steps

- Multi-Lingual, Multi-Domain Corpora
- Further explore the relationship to Quantitative Linguistics
- Model Linking Behavior

- Use it, for free (Apache 2.0 License)
  http://boilerpipe.googlecode.com