Characterizing Stylistic Elements in Syntactic Structure

Song Feng, Ritwik Banerjee, Yejin Choi

Stony Brook University
Emperor Constantine I supposedly transferred authority over Rome and the western part of the Roman Empire to the Pope by a decree.
The Beginning of Stylometry

Donation of Constantine

In 1439, Lorenzo Valla proved that it was a forgery, based on the comparison of the Latin used in this decree.

Lorenzo Valla
15th C.

21st C.

CFG Analysis
Outline

- Related work
- Sentence types
- Sentence outlines
- Tree topology
- Beyond production rules
- Experiments
- Conclusion
Outline

- Related work
  - Sentence types
  - Sentence outlines
  - Tree topology
  - Beyond production rules
  - Experiments
  - Conclusion
Why use deep syntax for authorship attribution?

- Rhetorical and compositional theories
  - Bain, 1887
  - Kemper, 1987
  - Strunk and White, 2008
Why use deep syntax for authorship attribution?

• Rhetorical and compositional theories
  ◦ Bain, 1887
  ◦ Kemper, 1987
  ◦ Strunk and White, 2008
Why use deep syntax for authorship attribution?

- Rhetorical and compositional theories
  - Bain, 1887
  - Kemper, 1987
  - Strunk and White, 2008

- Computational stylometric analysis and authorship attribution
  - Stamatatos et al., 2001
  - Baayen et al., 2002
  - Koppel and Schler, 2003

Deep syntactic elements
Why use deep syntax for authorship attribution?

- Rhetorical and compositional theories
  - Bain, 1887
  - Kemper, 1987
  - Strunk and White, 2008
- Computational stylometric analysis and authorship attribution
  - Stamatatos et al., 2001
  - Baayen et al., 2002
  - Koppel and Schler, 2003
PCFG models for stylometry

- Detecting distributional differences in sentence structures
  - Raghavan et al., 2010 (authorship attribution)
  - Sarawgi et al., 2011 (gender attribution)
  - Wong and Dras, 2011 (native language identification)
PCFG models for stylometry

- Detecting distributional differences in sentence structures
  - Raghavan et al., 2010 (authorship attribution)
  - Sarawgi et al., 2011 (gender attribution)
  - Wong and Dras, 2011 (native language identification)

But …
Short of providing clues about salient styles of sentence usage.
PCFG models for stylometry

- Detecting distributional differences in sentence structures
  ◦ Raghavan et al., 2010 (authorship attribution)
  ◦ Sarawgi et al., 2011 (gender attribution)
  ◦ Wong and Dras, 2011 (native language identification)

What are the **stylistic elements in sentence structures** that characterize individual authors?
Outline

- Related work
- Sentence types
- Sentence outlines
- Tree topology
- Beyond production rules
- Experiments
- Conclusion
“Christopher Columbus finally reached the shores of San Salvador after months of uncertainty at sea, the threat of mutiny, and a shortage of food and water.”

“After months of uncertainty at sea, the threat of mutiny, and a shortage of food and water, Christopher Columbus finally reached the shores of San Salvador.”
• “Christopher Columbus finally reached the shores of San Salvador after months of uncertainty at sea, the threat of mutiny, and a shortage of food and water.”

• “After months of uncertainty at sea, the threat of mutiny, and a shortage of food and water, Christopher Columbus finally reached the shores of San Salvador.”
• “Christopher Columbus finally reached the shores of San Salvador after months of uncertainty at sea, the threat of mutiny, and a shortage of food and water.”

• “After months of uncertainty at sea, the threat of mutiny, and a shortage of food and water, Christopher Columbus finally reached the shores of San Salvador.”
“Christopher Columbus finally reached the shores of San Salvador after months of uncertainty at sea, the threat of mutiny, and a shortage of food and water.”

- **Loose (cumulative)**

“After months of uncertainty at sea, the threat of mutiny, and a shortage of food and water, Christopher Columbus finally reached the shores of San Salvador.”
“Christopher Columbus finally reached the shores of San Salvador after months of uncertainty at sea, the threat of mutiny, and a shortage of food and water.”

- **Loose (cumulative)**

“After months of uncertainty at sea, the threat of mutiny, and a shortage of food and water, Christopher Columbus finally reached the shores of San Salvador.”
“Christopher Columbus finally reached the shores of San Salvador after months of uncertainty at sea, the threat of mutiny, and a shortage of food and water.”

- **Loose (cumulative)**

“After months of uncertainty at sea, the threat of mutiny, and a shortage of food and water, Christopher Columbus finally reached the shores of San Salvador.”

**Supporting clauses/phases**
“Christopher Columbus finally reached the shores of San Salvador after months of uncertainty at sea, the threat of mutiny, and a shortage of food and water.”

- **Loose (cumulative)**

“After months of uncertainty at sea, the threat of mutiny, and a shortage of food and water, Christopher Columbus finally reached the shores of San Salvador.”

**Supporting clauses/ phases**
Sentence Type - I

- “Christopher Columbus finally reached the shores of San Salvador after months of uncertainty at sea, the threat of mutiny, and a shortage of food and water.”
  - **Loose (cumulative)**

- “After months of uncertainty at sea, the threat of mutiny, and a shortage of food and water, Christopher Columbus finally reached the shores of San Salvador.”
“Christopher Columbus finally reached the shores of San Salvador after months of uncertainty at sea, the threat of mutiny, and a shortage of food and water.”

- **Loose (cumulative)**

“After months of uncertainty at sea, the threat of mutiny, and a shortage of food and water, Christopher Columbus finally reached the shores of San Salvador.”

- **Periodic**
Sentence Type Classification

- **Type-I classification**
  - Loose
  - Periodic

- **Type-II classification**
  - Simple
  - Complex
  - Compound
  - Complex-Compound
Sentence Type Classification

- Type-I classification
  - Loose
  - Periodic

- Type-II classification
  - Simple
  - Complex
  - Compound
  - Complex-Compound
Sentence Type Classification

- Type-I classification
  - Loose
  - Periodic

- Type-II classification
  - Simple
  - Complex
  - Compound
  - Complex-compound

Occurrence of main & supporting clauses
Sentence Type Classification

- Type-I classification
  - Loose
  - Periodic

- Type-II classification
  - Simple
  - Complex
  - Compound
  - Complex-compound

Occurrence of main & supporting clauses

Occurrence of independent & dependent clauses
# Type-II Classification

<table>
<thead>
<tr>
<th>Type</th>
<th>Sentence</th>
<th># ICs</th>
<th># DCs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simple</td>
<td>Jeju is a beautiful island.</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Complex</td>
<td>Jeju is so beautiful that we decided to stay for a few more days.</td>
<td>1</td>
<td>≥ 1</td>
</tr>
<tr>
<td>Compound</td>
<td>Jeju island is so beautiful and the food here is great too.</td>
<td>≥ 2</td>
<td>0</td>
</tr>
<tr>
<td>Complex-compound</td>
<td>Although I want to climb Halla, I haven't had the time, and haven't found anyone to go with.</td>
<td>≥ 2</td>
<td>≥ 1</td>
</tr>
</tbody>
</table>
# Type-II Classification

<table>
<thead>
<tr>
<th>Type</th>
<th>Sentence</th>
<th># ICs</th>
<th># DCs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simple</td>
<td>Jeju is a beautiful island.</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Complex</td>
<td>Jeju is so beautiful that we decided to stay for a few more days.</td>
<td>1</td>
<td>≥ 1</td>
</tr>
<tr>
<td>Compound</td>
<td>Jeju island is so beautiful and the food here is great too.</td>
<td>≥ 2</td>
<td>0</td>
</tr>
<tr>
<td>Complex-compound</td>
<td>Although I want to climb Halla, I haven't had the time, and haven't found anyone to go with.</td>
<td>≥ 2</td>
<td>≥ 1</td>
</tr>
</tbody>
</table>
## Type-II Classification

<table>
<thead>
<tr>
<th>Type</th>
<th>Sentence</th>
<th># ICs</th>
<th># DCs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simple</td>
<td>Jeju is a beautiful island.</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Complex</td>
<td>Jeju is so beautiful that we decided to stay for a few more days.</td>
<td>1</td>
<td>≥ 1</td>
</tr>
<tr>
<td>Compound</td>
<td>Jeju island is so beautiful and the food here is great too.</td>
<td>≥ 2</td>
<td>0</td>
</tr>
<tr>
<td>Complex-compound</td>
<td>Although I want to climb Halla, I haven't had the time, and haven't found anyone to go with.</td>
<td>≥ 2</td>
<td>≥ 1</td>
</tr>
</tbody>
</table>
# Type-II Classification

<table>
<thead>
<tr>
<th>Type</th>
<th>Sentence</th>
<th># ICs</th>
<th># DCs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simple</td>
<td>Jeju is a beautiful island.</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Complex</td>
<td>Jeju is so beautiful that we decided to stay for a few more days.</td>
<td>1</td>
<td>≥ 1</td>
</tr>
<tr>
<td>Compound</td>
<td>Jeju island is so beautiful and the food here is great too.</td>
<td>≥ 2</td>
<td>0</td>
</tr>
<tr>
<td>Complex-compound</td>
<td>Although I want to climb Halla, I haven't had the time, and haven't found anyone to go with.</td>
<td>≥ 2</td>
<td>≥ 1</td>
</tr>
</tbody>
</table>
## Type-II Classification

<table>
<thead>
<tr>
<th>Type</th>
<th>Sentence</th>
<th># ICs</th>
<th># DCs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simple</td>
<td>Jeju is a beautiful island.</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Complex</td>
<td>Jeju is so beautiful that we decided to stay for a few more days.</td>
<td>1</td>
<td>≥ 1</td>
</tr>
<tr>
<td>Compound</td>
<td>Jeju island is so beautiful and the food here is great too.</td>
<td>≥ 2</td>
<td>0</td>
</tr>
<tr>
<td>Complex-compound</td>
<td>Although I want to climb Halla, I haven't had the time, and haven't found anyone to go with.</td>
<td>≥ 2</td>
<td>≥ 1</td>
</tr>
</tbody>
</table>
Datasets

- **Scientific Papers**
  - ACL anthology reference corpus
    - Bird et al., 2008
  - 10 authors, 8 single-author papers per author

- **Novels**
  - 5 novelists
  - 5 novels for each author
  - First 3,000 sentences taken from each novel
Outline

• Related work
• Sentence types
• Sentence outlines
• Tree topology
• Beyond production rules
• Experiments
• Conclusion
Using parse trees to discover sentence outlines
Using parse trees to discover sentence outlines

Outline: \( S \rightarrow PP, VP \)
Comparing sentence outlines

<table>
<thead>
<tr>
<th>Hobbs</th>
<th>Joshi</th>
<th>Lin</th>
</tr>
</thead>
<tbody>
<tr>
<td>S (\rightarrow) S CC S.</td>
<td>S (\rightarrow) ADVP PP NP VP .</td>
<td>S (\rightarrow) SBAR NP VP .</td>
</tr>
<tr>
<td>S (\rightarrow) CC NP VP .</td>
<td>S (\rightarrow) PP NP ADVP VP .</td>
<td>FRAG (\rightarrow) NP : S .</td>
</tr>
<tr>
<td>S (\rightarrow) S VP .</td>
<td>S (\rightarrow) NP VP .</td>
<td>S (\rightarrow) NP VP .</td>
</tr>
<tr>
<td>S (\rightarrow) NP NP VP .</td>
<td>S (\rightarrow) S S CC S .</td>
<td>S (\rightarrow) PP VP .</td>
</tr>
<tr>
<td>S (\rightarrow) PP NP VP .</td>
<td>S (\rightarrow) ADVP NP VP .</td>
<td>S (\rightarrow) NP ADVP VP .</td>
</tr>
</tbody>
</table>
## Comparing sentence outlines

<table>
<thead>
<tr>
<th>Hobbs</th>
<th>Joshi</th>
<th>Lin</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S \rightarrow S \textbf{CC} S.$</td>
<td>$S \rightarrow \text{ADVP PP NP VP.}$</td>
<td>$S \rightarrow \text{SBAR NP VP.}$</td>
</tr>
<tr>
<td>$S \rightarrow \text{PP NP ADVP VP.}$</td>
<td>$S \rightarrow \text{NP VP.}$</td>
<td>$S \rightarrow \text{NP VP.}$</td>
</tr>
<tr>
<td>$S \rightarrow \text{NP NP VP.}$</td>
<td>$S \rightarrow \text{S S CC S.}$</td>
<td>$S \rightarrow \text{PP VP.}$</td>
</tr>
<tr>
<td>$S \rightarrow \text{PP NP VP.}$</td>
<td>$S \rightarrow \text{ADVP NP VP.}$</td>
<td>$S \rightarrow \text{NP ADVP VP.}$</td>
</tr>
</tbody>
</table>

**Compound sentences**
## Comparing sentence outlines

<table>
<thead>
<tr>
<th>Hobbs</th>
<th>Joshi</th>
<th>Lin</th>
</tr>
</thead>
<tbody>
<tr>
<td>( S \rightarrow S \text{ CC } S )</td>
<td>( S \rightarrow \text{ ADVP PP NP VP } ) ( S \rightarrow \text{ SBAR NP VP } )</td>
<td></td>
</tr>
<tr>
<td>( S \rightarrow \text{ NP VP } )</td>
<td>( S \rightarrow \text{ NP VP } ) ( S \rightarrow \text{ NP VP } )</td>
<td>( S \rightarrow \text{ NP VP } ) ( S \rightarrow \text{ PP VP } )</td>
</tr>
<tr>
<td>( S \rightarrow \text{ NP VP } )</td>
<td>( S \rightarrow \text{ S S CC } S ) ( S \rightarrow \text{ PP VP } )</td>
<td>( S \rightarrow \text{ NP VP } ) ( S \rightarrow \text{ NP ADVP VP } )</td>
</tr>
</tbody>
</table>

### Starting with adverbial clauses
Comparing sentence outlines

<table>
<thead>
<tr>
<th>Hobbs</th>
<th>Joshi</th>
<th>Lin</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S \rightarrow S \text{ CC } S.$</td>
<td>$S \rightarrow \text{ ADVP PP NP VP}$</td>
<td>$S \rightarrow \text{ SBAR NP VP}$</td>
</tr>
<tr>
<td>$S \rightarrow \text{ CC NP VP}$</td>
<td></td>
<td>FRAG $\rightarrow \text{ NP : S}$</td>
</tr>
<tr>
<td>$S \rightarrow S \text{ VP}$</td>
<td></td>
<td>$S \rightarrow \text{ NP VP}$</td>
</tr>
<tr>
<td>$S \rightarrow \text{ NP NP VP}$</td>
<td>$S \rightarrow S S \text{ CC } S.$</td>
<td>$S \rightarrow \text{ PP VP}$</td>
</tr>
<tr>
<td>$S \rightarrow \text{ PP NP VP}$</td>
<td>$S \rightarrow \text{ ADVP NP VP}$</td>
<td>$S \rightarrow \text{ NP ADVP VP}$</td>
</tr>
</tbody>
</table>

Complex sentences
Outline

- Related work
- Sentence types
- Sentence outlines
- Tree topology
- Beyond production rules
- Experiments
- Conclusion
“For processing free texts, hand-crafted grammars are neither practical nor reliable.”

“These algorithms cannot deal with words for which classifiers have not been learned.”
“For processing free texts, hand-crafted grammars are neither practical nor reliable.”

“These algorithms cannot deal with words for which classifiers have not been learned.”
“For processing free texts, hand-crafted grammars are neither practical nor reliable.”

“These algorithms cannot deal with words for which classifiers have not been learned.”
“For processing free texts, hand-crafted grammars are neither practical nor reliable.”

“These algorithms cannot deal with words for which classifiers have not been learned.”
“For processing free texts, hand-crafted grammars are neither practical nor reliable.”

“TheSE algorithms cannot deal with words for which classifiers have not been learned.”
Tree topology: metrics

- Leaf height
- Furcation height
- Level width
- Horizontal imbalance
- Vertical imbalance
Tree topology: leaf height

Leaf height ("texts") = 6
Tree topology: furcation height

Furcation height \((VP_2) = 3\)
Tree topology: **level width**

**Level Width**(level$_3$) = 8
Tree topology: imbalance

Horizontal Imbalance (PP)
= |width(IN) – width(S₂)|
= |1 – 3| = 2

Vertical Imbalance (PP)
= |height(IN) – height(S₂)|
= |2 – 6| = 4
Tree topology metrics: novelists

<table>
<thead>
<tr>
<th>Tree-topology metrics</th>
<th>Charles Dickens</th>
<th>Edward Bulwer-Lytton</th>
<th>Jane Austen</th>
<th>Thomas Hardy</th>
<th>Walter Scott</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sentence Length</td>
<td>24.1</td>
<td>26.7</td>
<td>31.4</td>
<td>21.5</td>
<td>34.1</td>
</tr>
<tr>
<td>Leaf Height</td>
<td>4.7</td>
<td>5.0</td>
<td>5.4</td>
<td>4.9</td>
<td>5.9</td>
</tr>
<tr>
<td>Furcation Height</td>
<td>1.9</td>
<td>1.9</td>
<td>2.1</td>
<td>1.9</td>
<td>2.1</td>
</tr>
<tr>
<td>Level Width</td>
<td>4.1</td>
<td>4.4</td>
<td>4.7</td>
<td>3.8</td>
<td>4.9</td>
</tr>
<tr>
<td>Imbalance</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Horizontal</td>
<td>1.1</td>
<td>1.1</td>
<td>1.3</td>
<td>1.2</td>
<td>1.4</td>
</tr>
<tr>
<td>Vertical</td>
<td>1.0</td>
<td>1.1</td>
<td>1.2</td>
<td>1.0</td>
<td>1.4</td>
</tr>
</tbody>
</table>
# Tree topology metrics: novelists

<table>
<thead>
<tr>
<th>Tree-topology metrics</th>
<th>Charles Dickens</th>
<th>Edward Bulwer-Lytton</th>
<th>Jane Austen</th>
<th>Thomas Hardy</th>
<th>Walter Scott</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sentence Length</td>
<td>24.1</td>
<td>26.7</td>
<td>31.4</td>
<td>21.5</td>
<td>34.1</td>
</tr>
<tr>
<td>Leaf Height</td>
<td>4.7</td>
<td>5.0</td>
<td>5.4</td>
<td>4.9</td>
<td>5.9</td>
</tr>
<tr>
<td>Furcation Height</td>
<td>1.9</td>
<td>1.9</td>
<td>2.1</td>
<td>1.9</td>
<td>2.1</td>
</tr>
<tr>
<td>Level Width</td>
<td>4.1</td>
<td>4.4</td>
<td>4.7</td>
<td>3.8</td>
<td>4.9</td>
</tr>
<tr>
<td>Imbalance</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Horizontal</td>
<td>1.1</td>
<td>1.1</td>
<td>1.3</td>
<td>1.2</td>
<td>1.4</td>
</tr>
<tr>
<td>Vertical</td>
<td>1.0</td>
<td>1.1</td>
<td>1.2</td>
<td>1.0</td>
<td>1.4</td>
</tr>
</tbody>
</table>
# Tree topology metrics: novelists

<table>
<thead>
<tr>
<th>Tree-topology metrics</th>
<th>Charles Dickens</th>
<th>Edward Bulwer-Lytton</th>
<th>Jane Austen</th>
<th>Thomas Hardy</th>
<th>Walter Scott</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sentence Length</td>
<td>24.1</td>
<td>26.7</td>
<td>31.4</td>
<td>21.5</td>
<td>34.1</td>
</tr>
<tr>
<td>Leaf Height</td>
<td>4.7</td>
<td>5.0</td>
<td>5.4</td>
<td>4.9</td>
<td>5.9</td>
</tr>
<tr>
<td>Furcation Height</td>
<td>1.9</td>
<td>1.9</td>
<td>2.1</td>
<td>1.9</td>
<td>2.1</td>
</tr>
<tr>
<td>Level Width</td>
<td>4.1</td>
<td>4.4</td>
<td>4.7</td>
<td>3.8</td>
<td>4.9</td>
</tr>
<tr>
<td>Imbalance</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Horizontal</td>
<td>1.1</td>
<td>1.1</td>
<td>1.3</td>
<td>1.2</td>
<td>1.4</td>
</tr>
<tr>
<td>Vertical</td>
<td>1.0</td>
<td>1.1</td>
<td>1.2</td>
<td>1.0</td>
<td>1.4</td>
</tr>
</tbody>
</table>
# Tree topology metrics: novelists

<table>
<thead>
<tr>
<th>Tree-topology metrics</th>
<th>Novels</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Charles Dickens</td>
</tr>
<tr>
<td>Sentence Length</td>
<td>24.1</td>
</tr>
<tr>
<td>Leaf Height</td>
<td>4.7</td>
</tr>
<tr>
<td>Furcation Height</td>
<td>1.9</td>
</tr>
<tr>
<td>Level Width</td>
<td>4.1</td>
</tr>
<tr>
<td>Imbalance</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Horizontal</td>
</tr>
<tr>
<td></td>
<td>Vertical</td>
</tr>
</tbody>
</table>
# Tree topology metrics: novelists

<table>
<thead>
<tr>
<th>Tree-topology metrics</th>
<th>Novels</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Charles Dickens</td>
</tr>
<tr>
<td>Sentence Length</td>
<td>24.1</td>
</tr>
<tr>
<td>Leaf Height</td>
<td>4.7</td>
</tr>
<tr>
<td>Furcation Height</td>
<td>1.9</td>
</tr>
<tr>
<td>Level Width</td>
<td>4.1</td>
</tr>
</tbody>
</table>

**Periodic**

**Loose**
Outline

- Related work
- Sentence types
- Sentence outlines
- Tree topology
- Beyond production rules
- Experiments
- Conclusion
PCFG: production rules

Pr: VP₁ \rightarrow VBG NP₁
Beyond PCFG production rules

Pr*: \( VP_1 \uparrow ^* S_2 \rightarrow VBG \ NP_1 \)
Beyond PCFG production rules

Pr*: NNS₁ → “texts”

Pr^*: NNS₁ ^ NP₁ → “texts”
Beyond PCFG production rules

\[
\text{Syn}^\uparrow: \ VP_1 \rightarrow S \rightarrow PP
\]
Beyond PCFG production rules

\[
S_1 
\rightarrow \quad S_2 
\rightarrow \quad PP 
\rightarrow \quad IN 
\rightarrow \quad For 
\rightarrow \quad VBG 
\rightarrow \quad VP_1 
\rightarrow \quad NP_1 
\rightarrow \quad JJ_1 
\rightarrow \quad free 
\rightarrow \quad NNS_1 
\rightarrow \quad texts 
\rightarrow \quad VP_2 
\rightarrow \quad NNS_2 
\rightarrow \quad VBP 
\rightarrow \quad ADJP 
\rightarrow \quad JJ_3 
\rightarrow \quad CC 
\rightarrow \quad JJ_4 
\rightarrow \quad DT 
\rightarrow \quad neither 
\rightarrow \quad practical 
\rightarrow \quad nor 
\rightarrow \quad reliable 
\rightarrow \quad hand-crafted 
\rightarrow \quad grammars 
\rightarrow \quad are 
\rightarrow \quad Syn \downarrow : VP_1 \rightarrow VBG , VP_1 \rightarrow NP_1
\]
Outline

- Related work
- Sentence types
- Sentence outlines
- Tree topology
- Beyond production rules
- Experiments
- Conclusion
Experiments

- SVM classifier (LIBLINEAR)
- 5-fold cross validation
  - 80% training, 20% testing
  - 20% training, 80% testing
Experiments

- SVM classifier (LIBLINEAR)
- 5-fold cross validation
  - 80% training, 20% testing
  - 20% training, 80% testing

Sufficient training data may not be available in practical scenarios (e.g., forensics). (Luyckx and Daelemans, 2008)
Experiments

- SVM classifier (LIBLINEAR)
- 5-fold cross validation
  - 80% training, 20% testing
  - 20% training, 80% testing
- Features
  - PCFG rule-based
  - STYLE_{11}
    - 6 parameters from distribution of sentence types
    - 5 topological metrics
Experiments

% of sentences that are
1. Simple
2. Complex
3. Compound
4. Complex-compound
5. Loose
6. Periodic

- STYLE$_{11}$
  - 6 parameters from distribution of sentence types
  - 5 topological metrics
Experiments

- SVM classifiers built using LibLinear
- 5-fold cross validation
- 80% training, 20% testing

- STYLE
  - 6 parameters from distribution of sentence types
  - 5 topological metrics

1. Leaf height
2. Furcation height
3. Level-width
4. Horizontal imbalance
5. Vertical imbalance
Experimental results

Scientific Papers: 20% training data

Parse-tree features
Experimental results

Scientific Papers: 20% training data

- **unigrams**
- **pr^***
- **syn↑***
- **syn*v+h**

- Parse-tree features
- Parse-tree + Style11 features
Scientific Papers: 20% training data

Best unlexicalized feature ($pr^\wedge$): 60.6%
Experimental results

Novels: 20% training data

- unigrams
- pr^*
- syn↑*
- syn*v+h

- Parse-tree features
- Parse-tree + Style11 features
Experimental results

Novels: 20% training data

Best unlexicalized feature (syn_{v+h}): 73.2%
Unlexicalized features across domains

Training v/s Performance: unlexicalized features

Scientific Papers: 32.9% trained on 20% data, 17.0% trained on 80% data

Novels: 17.0% trained on 20% data, 32.9% trained on 80% data
Conclusions

- Analyzed writing styles with *interpretable* characterization of stylistic elements.
- Even without lexical elements, features derived from *sentence structures* alone can predict authorship with high accuracy.
- Using *topological features of parse trees* in conjunction with features derived from production rules provide the best results in authorship attribution.
- Even with little training data, our techniques provide reasonably good performance.