Analysis on English Text Genre Classification Based on Dependency Types

QUALICO 2018 July 5-8, Wrocław, Poland

Yaqin Wang, Zhejiang University, wyaqin@zju.edu.cn
Outline

- Introduction
- Materials and Methods
- Results and Discussion
- Conclusion
Introduction

1. Literature Review
2. Research Questions
Literature Review

- Methods employed in text genre classification:
  
  Lexical features are the most common methods. Syntactic information is also important.

- Less sufficient explicit attention to the syntactic features in genre classification.

(Baayen et al., 1996; Stamatatos et al., 2000a, 2000b; Feldman et al., 2009)
Literature Review

- **Dependency grammar** describes the unequal syntactic relationship between the words concerned.

- Dependency type can well capture the syntactic relationship between words.
Literature Review

- Several studies using syntactic annotations under the framework of dependency grammar have been conducted in text genre classification and authorship verification.

- Little consideration has been given to the role that the dependency type play in classifying English genres and texts.

(Hollingsworth, 2012; Gao and Feng, 2011; Rygi, 2014)
Research Questions

- What is the distribution of dependency types in different genres of English?

- Can dependency type be a useful metric in text clustering?

- If so, do all dependency types play the same role in classifying texts? (If not, why?)
Materials and Methods

1. Treebank establishment
2. Methods
Treebank Establishment

- British National Corpus (BNC)

5 texts per genre
4 genres
20 texts in total

Informative

Imaginative
Written to be Spoken
Spoken

Applied Science
Arts
Belief
Commerce
Leisure
Natural Science
Social Science
World Affairs

Materials and Methods
Treebank Establishment

- Stanford parser (version 3.7.0)

  det (boy-2, The-1)
  nsubj (has-3, boy-2)
  root (ROOT-0, has-3)
  det (apple-5, an-4)
  dobj (has-3, apple-5)
Methods

- Principal component analysis (PCA)
- Agglomerative clustering (Cophenetic Correlation Coefficient (CPCC))
  To see the clustering result and the relationship between texts.
- Random Forest
  To determine whether each dependency type is important for text classification.
Results and Discussion

1. The distribution of dependency types
2. Clustering results
3. The importance of dependency types
The distribution of dependency types
The distribution of dependency types

<table>
<thead>
<tr>
<th>TYPE</th>
<th>IN</th>
<th>IM</th>
<th>WS</th>
<th>SP</th>
</tr>
</thead>
<tbody>
<tr>
<td>nsubj</td>
<td>0.061</td>
<td>0.097</td>
<td>0.086</td>
<td>0.111</td>
</tr>
<tr>
<td>nmod</td>
<td>0.108</td>
<td>0.079</td>
<td>0.099</td>
<td>0.065</td>
</tr>
<tr>
<td>advmod</td>
<td>0.036</td>
<td>0.049</td>
<td>0.037</td>
<td>0.060</td>
</tr>
<tr>
<td>amod</td>
<td>0.069</td>
<td>0.040</td>
<td>0.042</td>
<td>0.032</td>
</tr>
<tr>
<td>ccomp</td>
<td>0.009</td>
<td>0.019</td>
<td>0.018</td>
<td>0.028</td>
</tr>
</tbody>
</table>

- **ccomp**: the clausal complement of a verb or adjective is a dependent clause which is a core argument. “He says that you like to swim” ccomp(says, like)

- The frequency of ccomp in *spoken* texts’ dependency type distribution is greater than which in *informative* texts. Spoken genre tends to have more clausal complements (finite or non-finite) than do written texts.

Results and Discussion

QUALICO 2018
The distribution of dependency types

- Differences of dependency types across genres do exist.
- This distribution may be consistent with stylistic transformation from spoken texts to written texts.
**Clustering results**

- **PCA**

<table>
<thead>
<tr>
<th></th>
<th>PC1</th>
<th>PC2</th>
<th>PC3</th>
<th>PC4</th>
<th>PC5</th>
<th>PC6</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Standard deviation</strong></td>
<td>3.65</td>
<td>2.27</td>
<td>2.19</td>
<td>1.89</td>
<td>1.41</td>
<td>1.34</td>
</tr>
<tr>
<td><strong>Proportion of Variance</strong></td>
<td>0.34</td>
<td>0.13</td>
<td>0.12</td>
<td>0.09</td>
<td>0.05</td>
<td>0.05</td>
</tr>
<tr>
<td><strong>Cumulative Proportion</strong></td>
<td>0.34</td>
<td>0.47</td>
<td><strong>0.60</strong></td>
<td>0.69</td>
<td>0.74</td>
<td>0.78</td>
</tr>
</tbody>
</table>

Results and Discussion

---

2018 QUALICO
Clustering results

- Agglomerative Clustering

Cluster Dendrogram

Informative & Written to be spoken

Spoken & Imaginative

CCPC = 0.72
Clustering results

- Agglomerative Clustering

- Consistent with PCA results.
- Dependency type may be a useful classifying method.
Clustering results

- Spoken English is isolated from informative and written to be spoken genre.
- Although imaginative genre belongs to written texts, they share more characteristics with spoken texts. (large amount of conversations in imaginative texts.)
- Written to spoken genre, though written generally for spoken purpose, is much closer to written genres.
Random Forests

More important: ccomp, nsubj, dep, cop
Less important: csubjpass, parataxis, nmod:tmod

Results and Discussion
Random Forests

- Ten most important types (in terms of mean decrease accuracy):
  ccomp, nsubj, dep, cop, advmod, acl:relcl, xcomp, advcl, root, dobj
- The CPCC of the dendrogram and distance matrix of the texts is 0.75(0.72).
The distribution of dependency types

<table>
<thead>
<tr>
<th>TYPE</th>
<th>IN</th>
<th>IM</th>
<th>WS</th>
<th>SP</th>
</tr>
</thead>
<tbody>
<tr>
<td>nsubj</td>
<td>0.061</td>
<td>0.097</td>
<td>0.086</td>
<td>0.111</td>
</tr>
<tr>
<td>punct</td>
<td>0.101</td>
<td>0.142</td>
<td>0.100</td>
<td>0.094</td>
</tr>
<tr>
<td>case</td>
<td>0.118</td>
<td>0.088</td>
<td>0.112</td>
<td>0.076</td>
</tr>
<tr>
<td>det</td>
<td>0.099</td>
<td>0.086</td>
<td>0.087</td>
<td>0.070</td>
</tr>
<tr>
<td>root</td>
<td>0.040</td>
<td>0.053</td>
<td>0.054</td>
<td>0.065</td>
</tr>
<tr>
<td>nmod</td>
<td>0.108</td>
<td>0.079</td>
<td>0.099</td>
<td>0.065</td>
</tr>
<tr>
<td>advmod</td>
<td>0.036</td>
<td>0.049</td>
<td>0.037</td>
<td>0.060</td>
</tr>
<tr>
<td>dobj</td>
<td>0.038</td>
<td>0.046</td>
<td>0.039</td>
<td>0.047</td>
</tr>
<tr>
<td>mark</td>
<td>0.032</td>
<td>0.036</td>
<td>0.032</td>
<td>0.044</td>
</tr>
<tr>
<td>aux</td>
<td>0.024</td>
<td>0.031</td>
<td>0.029</td>
<td>0.035</td>
</tr>
<tr>
<td>cc</td>
<td>0.034</td>
<td>0.028</td>
<td>0.034</td>
<td>0.033</td>
</tr>
<tr>
<td>amod</td>
<td>0.069</td>
<td>0.040</td>
<td>0.042</td>
<td>0.032</td>
</tr>
<tr>
<td>cop</td>
<td>0.017</td>
<td>0.017</td>
<td>0.019</td>
<td>0.029</td>
</tr>
<tr>
<td>ccomp</td>
<td>0.009</td>
<td>0.019</td>
<td>0.018</td>
<td>0.028</td>
</tr>
<tr>
<td>compound</td>
<td>0.040</td>
<td>0.016</td>
<td>0.039</td>
<td>0.026</td>
</tr>
<tr>
<td>conj</td>
<td>0.035</td>
<td>0.030</td>
<td>0.031</td>
<td>0.025</td>
</tr>
<tr>
<td>dep</td>
<td>0.012</td>
<td>0.013</td>
<td>0.011</td>
<td>0.024</td>
</tr>
<tr>
<td>xcomp</td>
<td>0.013</td>
<td>0.018</td>
<td>0.017</td>
<td>0.021</td>
</tr>
</tbody>
</table>

Further research (e.g., larger amounts of texts and more numbers of genres) is needed...

Results and Discussion
Conclusions

- Differences across genres exist due to different genre features.
- Dependency types can be used as an effective parameter in distinguishing text genres, especially between *spoken* genre and *written* genre.
- Some dependency types are more important than others in classifying genres.
Future Directions

- More numbers of texts from more genres and more languages.
- Combining other lexical and syntactic features.
- Theoretical explanations are needed.
References


Thank you!

QUALICO 2018 July 5-8, Wrocław, Poland

Yaqin Wang, Zhejiang University, wyaqin@zju.edu.cn