Automatically Distinguishing between Written Output Produced by Heritage and Non-Heritage Learners of Polish as a Foreign Language

Simon Zuberek
December 13th, 2022
Heritage Language Learners

Students who have been raised with a strong cultural connection to a particular language through family interaction¹.

Basic Skills Difficulty
Non-Heritage v. Heritage Learners

<table>
<thead>
<tr>
<th>Non-Heritage</th>
<th>Heritage</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Speaking (Productive)</td>
<td>1. Writing (Productive)</td>
</tr>
<tr>
<td>2. Writing (Productive)</td>
<td>2. Reading (Receptive)</td>
</tr>
<tr>
<td>3. Listening (Receptive)</td>
<td>3. Speaking (Productive)</td>
</tr>
<tr>
<td>4. Reading (Receptive)</td>
<td>4. Listening (Receptive)</td>
</tr>
</tbody>
</table>

Where 1 is most difficult and 4 is least difficult
Why is writing hard?

- Polish nominal morphology (i.e. cases)
- Nominal mistakes made by heritage learners of Polish\(^2\):
  - Overgeneralize the use of the LOC case after certain bivalent prepositions.
  - Express DOs in the ACC case following the verbs requiring GEN DOs.
  - Express DOs in the ACC case following negated verbs (normally requiring GEN)

Research Questions

1. Are these error types enough to differentiate between non-heritage and heritage written output? (Relevant for software design)

2. Are some of these error types more consequential than others in making that distinction? (Relevant for instructional design)
Approach

Develop a supervised ML classifier that distinguishes between heritage and non-heritage written output based on the errors committed.

Features:
- Counts of LOC pos-prepositionally
- Counts of GEN objects following verbs that take the GEN case
- Counts of GEN objects following negated verbs
- Per-character entropy
## Data

### Training
- The corpus of [Heritage Language Variation and Change (HLVC)](Heritage%20Language%20Variation%20and%20Change%20(HLVC))
- Interviews with Polish heritage speakers
- [PoLKo, the Polish Learner Corpus](PoLKo%20the%20Polish%20Learner%20Corpus)

### Testing
- Essays written by non-heritage learners of Polish as a foreign language (UIC)
- Essays written by heritage learners of Polish as a foreign language (UIC)
Data (continued)

Training
- 41 Heritage Docs:
  - HLVC: 35 interviews
  - In-person: 6 interviews
- 36 Non-Heritage Docs:
  - PoLKo: 36 essays

Testing
- 9 heritage essays
- 9 non-heritage essays

Baseline: 0.5
## Results: Test Data

### Test Data Accuracy:

<table>
<thead>
<tr>
<th>Model</th>
<th>LOC after Prepositions</th>
<th>Genitive after Verbs</th>
<th>Genitive of Negation</th>
<th>Per-Character Entropy</th>
<th>All Features Combined</th>
</tr>
</thead>
<tbody>
<tr>
<td>MultinomialNB(alpha=1)</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>ComplementNB()</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>DecisionTreeClassifier()</td>
<td>0.389</td>
<td>0.556</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>RandomForestClassifier()</td>
<td>0.389</td>
<td>0.556</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>SVC(kernel='linear')</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
</tr>
</tbody>
</table>
## Results: Cross-Validated Train Data

### Cross-Validated Training Data Accuracy:

<table>
<thead>
<tr>
<th>Model</th>
<th>LOC after Prepositions</th>
<th>Genitive after Verbs</th>
<th>Genitive of Negation</th>
<th>Per-Character Entropy</th>
<th>All Features Combined</th>
</tr>
</thead>
<tbody>
<tr>
<td>MultinomialNB()</td>
<td>0.489</td>
<td>0.489</td>
<td>0.489</td>
<td>0.489</td>
<td>0.668</td>
</tr>
<tr>
<td>ComplementNB()</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
<td>0.668</td>
</tr>
<tr>
<td>DecisionTreeClassifier()</td>
<td>0.853</td>
<td>0.794</td>
<td>0.839</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>RandomForestClassifier()</td>
<td>0.901</td>
<td>0.744</td>
<td>0.826</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>SVC()</td>
<td>0.864</td>
<td>0.622</td>
<td>0.596</td>
<td></td>
<td>0.864</td>
</tr>
</tbody>
</table>
Discussion

Challenges
- Data scarcity
- Difference in text genres (training vs. testing)

Test Data
- Same as the baseline
- Tree-based algorithms
- Feature Ranking:
  1. GEN post Vs
  2. Negation
  3. LOC post PPs

Cross-Validated
- Better than the baseline
- Tree-based algorithms
- Feature Ranking:
  1. LOC post PPs
  2. Negation
  3. GEN post Vs
Research Questions Revisited

1. Are these error types enough to differentiate between non-heritage and heritage learner output?

A cautious “yes”, given more data representing the same genre.

2. Are some of these error types more consequential than others in making that distinction?

(Relevant for instructional design)

A definitive “yes”, with concrete pedagogical ramifications.
Thank you.