Grammatical Accents: Using Machine Learning to Quantify Language Transfer

Tiwalayo Eisape, William Merrill, Sven Dietz, Joshua K. Hartshorne

1Department of Psychology, Boston College; 2Department of Linguistics, Yale University

Background

Native-language identification has been proven possible when a wide set of features is applied to the task [1]. Further more, languages besides English have been widely ignored (Q3).

As a first step, we broaden our language set to include Spanish while simultaneously restricting our feature set to exclusively include Spanish while simultaneously accounting for the variable length of essays.

**Features**

- **Tree kernels:** Clustered representations of syntactic trees
- **Dependency parsed representations of sentences**

**Models / Results**

- **Support Vector Machines (SVM) and Feed Forward Neural Networks (FFN)** are clustered features.

**Conclusions / Future Directions**

- **Next Steps:**
  1. Expand features to further encapsulate syntax
  2. Open up the black box.
  3. Reverse engineer our learning algorithms for interpretation

**Acknowledgements**

Special thanks to William Merrill, Clinton Tak, and the rest of the Language Learning Lab. TE is supported by the Ronald E. McNair Scholarship (TRIO) and JKH is supported by the Academic Technology Innovation Grant (Boston College)

**REFERENCES**


**Insights**

1. Spanish and Italian - same language family. Italio-Western Romance
2. Hindi and Telugu - high proximity and language sharing
3. Arabic and Chinese
4. English
5. French
6. German
7. Italian
8. Japanese
9. Korean
10. Spanish
11. Turkish

**Conclusions**

- By achieving state of the art accuracy, using strictly syntactic features, we show machine learning can pick up on generalizable, grammatical idiosyncrasies associated with (L1->L2) language transfer.

**Insights**

- SVM - TOEFL
  - Confusion matrix - without reduction
  - Arabi: 29; Chinese: 27; French: 28; German: 5; Italian: 3; Japanese: 1; Korean: 0; Portuguese: 2; Russian: 1; Spanish: 2; Turkish: 0

**Models / Results**

- **Support Vector Machines (SVM) and Feed Forward Neural Networks (FFN)** are clustered features.

**Language Learning**

- TOEFL
  - Arabic
  - Chinese
  - French
  - German
  - Italian
  - Japanese
  - Korean
  - Portuguese
  - Russian
  - Spanish

**Conclusions / Future Directions**

- By achieving state of the art accuracy, using strictly syntactic features, we show machine learning can pick up on generalizable, grammatical idiosyncrasies associated with (L1->L2) language transfer.

**Next Steps:**

1. Expand features to further encapsulate syntax
2. Open up the black box.
3. Reverse engineer our learning algorithms for interpretation