Grammatical Accents: Using Machine Learning to Quantify Language Transfer

Tiwalayo Eisape[1], William Merrill[2], Sven Dietz[1], Joshua K. Hartshorne[1]
[1]Department of Psychology Boston College, [2]Department of Linguistics Yale University

Language learning lab

Motivation

Goal :
To use machine learning to establish a broad-based method to empirically study the effects of first language syntax on second language (L1->L2 transfer).

Q_1: Does NLI work in languages other than english [cf. 1]?

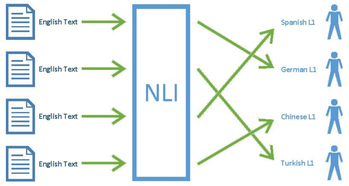
Q_{2a}: What grammatical features can we train on successfully? Which are the most informative?

Q_{2b}: Which are the most accurate* classifiers

Q_3: Can machine learning algorithms learn L1->L2 patterns that generalize across L2s?

Q_4: Are only certain parts of input (i.e language) informative? Which ones? [7]

Native-Language Identification (NLI):
The process of determining an author's native language (L1) based only on their writings in a second language (L2)



(Malmasi and Dras 2017)

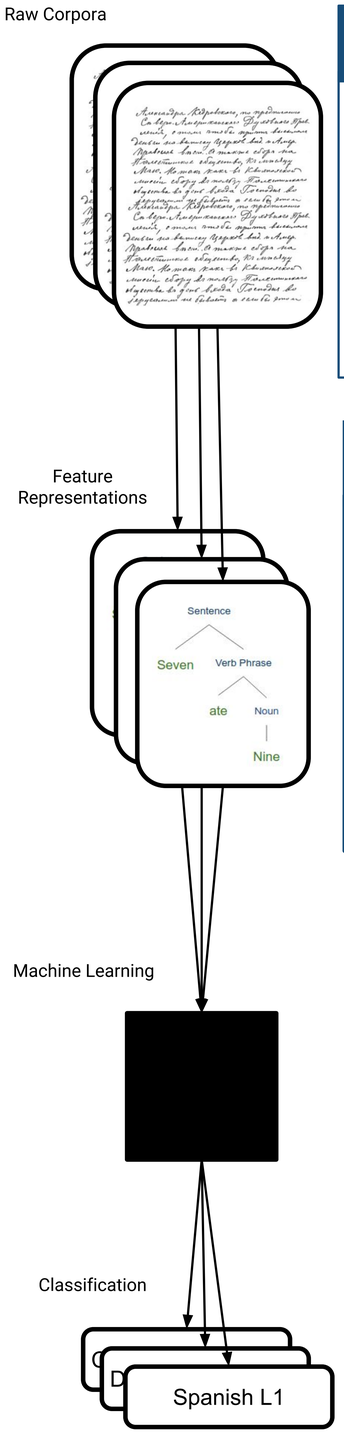
Method:
Compare the results of a variety of state-of-the-art machine learning techniques on NLI in two languages: English and Spanish.

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 (Q_1). As a first step, we broaden our language set to include Spanish while simultaneously restricting our feature set to exclusively syntactic features as inspired by [4].

- [3] POS n-grams <= 4-grams, dependency labels.
- [4] POS n-grams <= 4-grams. Used SVMs and shallow neural networks, achieving accuracy > 50%.
- [5] POS n-grams <= tri-grams. Used SVMs to achieve accuracy > 50%

*Accuracy was measured as a weighted average of the F1 scores of each class where F1 = 2 • (Recall • Precision) / (Recall + Precision)



Corpora

Q_1

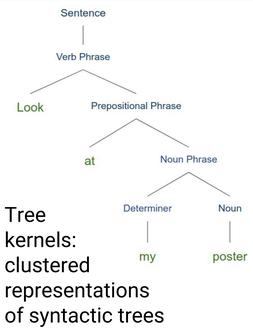
We perform NLI on datasets in two languages: English (TOEFL) and Spanish (CAES)

TOEFL

Instituto Cervantes

Spanish
Hindi Italian Korean Telugu
Turkish German Japanese
Arabic Chinese French
Russian Portuguese
English

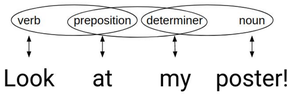
Features

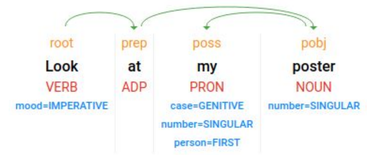


Tree kernels: clustered representations of syntactic trees

Q_{2a} Features used include labeled and unlabeled tree kernels as well as part of speech and dependency tags (TF-IDF weighting [4] was used to emphasize infrequency). Tags were generated using SyntaxNet [8].

Part of speech tags
n-grams up to
and including tri-grams





Dependency parsed representations of sentences

Models / Results

	POS	Labels	POS + Labels	
CAES	52.06	50.08	52.7	FF
	61.93	53.93	55.95	SVM
TOEFL	28.25	29.25	37.78	FF
	45.72	43.68	51.47	SVM

Support Vector Machines (SVM) and Feed Forward Neural Networks (FF) are fed clustered features.

Recurrent Neural Networks (RNN) and Convolutional Neural Networks (CNN) were fed features serially using padding to account for the variable length of essays

$Q_{2a} + Q_{2b}$

	POS			Labels			
	50	100	150	50	100	150	
CAES	46.35	48.6	50.25	40.38	49.5	41.28	CNN
	44.4	51.7	43.1	42.4	44.51	42.68	RNN
TOEFL	18.65	21.29	25.69	19.69	25.33	23.47	CNN
	14.43	18.27	22.9	16.2	20.2	21.82	RNN

chance for CAES = 16.67%, chance for TOEFL = 9.09%

TOEFL 52.6%

Q_3

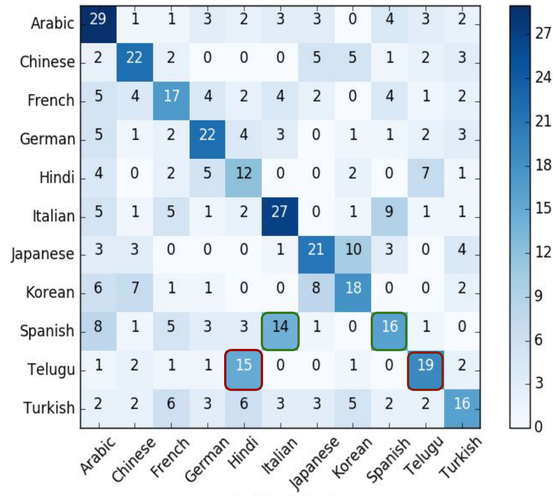
Cross-validating across languages shows some aspects of transfer generalize beyond individual L2s

Cross-validation performed on intersection of languages using FF, chance = 33.33%

	Labeled Tree Kernels	Unlabeled Tree Kernels
CAES	25.5	8.54
TOEFL	42.98	54.76

Labeled and unlabeled tree kernels [6] represent syntactic trees both by structure alone and by structure coupled with dependency labels

Insights



Insights:
1. Spanish and Italian - same language family: Italo-Western Romance
2. Hindi and Telugu - high proximity and language sharing

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 "Super Tagging" [2]
2. Open up the black box.
   Q_4 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)

TRIO Ronald E. McNair Post-Baccalaureate Achievement Program

REFERENCES:
[1] Tetreault, Joel, Daniel Blanchard, and Aoife Cahill. "A report on the first native language identification shared task." Proceedings of the eighth workshop on innovative use of NLP for building educational applications. 2013.
[2] Joshi, A. K. and Srinivas, B. Disambiguation of Super Parts of Speech (or Supertags): Almost Parsing. Proceedings of the 17th International Conference on Computational Linguistics Kyoto, Japan. 1994.
[3] Berzak, Yevgeni, Roi Reichart, and Boris Katz. "Reconstructing native language typology from foreign language usage." arXiv preprint arXiv:1404.6312 (2014).
[4] Gebre, Binyam Gebrekidan, et al. "Improving native language identification with tf-idf weighting." the 8th NAACL Workshop on Innovative Use of NLP for Building Educational Applications (BEA8). 2013.
[5] Jarvis, Scott, Yves Bestgen, and Steve Pepper. "Maximizing classification accuracy in native language identification." Proceedings of the eighth workshop on innovative use of NLP for building educational applications. 2013.
[6] Kruengkrai, Canasai, et al. "Language identification based on string kernels." Communications and Information Technology, 2005. ISCIT 2005. IEEE International Symposium on. Vol. 2. IEEE, 2005.
[7] Johnson, Jacqueline S., and Elissa L. Newport. "Critical period effects on universal properties of language: The status of subjacency in the acquisition of a second language." Cognition 39.3 (1991): 215-258.
[8] Petrov, Slav. "Announcing syntaxnet: The world's most accurate parser goes open source." Google Research Blog 12 (2016).