# What's in a Word: Classification of English-Translated Texts by Source Language

*Eunji Lee, Katherine Wu*

CS221: Artificial Intelligence: Principles and Techniques, Stanford University

Mentor: Jerry Qu

Stanford Computer Science

## Motivation

- English is a "**universal language**"
- Sometimes **source language is unknown**:
  - Places **cultural context** to literature
- **Develop** and **evaluate** effectiveness of **machine translation models.**
- **3+ language classification** rarely attempted

## Problem Definition

**Goal:**

Predict **source language** of a text translated to English

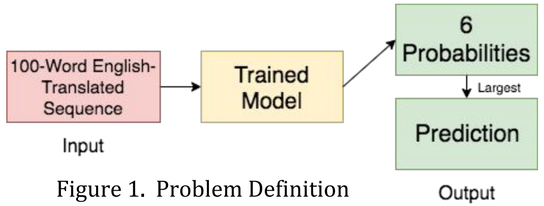


Figure 1. Problem Definition

**Evaluation:** Percentage of texts correctly classified into original language

## Challenges

- Data collection and preprocessing:
  - − **Author/translator** signals and noise
  - \+ **Vary texts**' authors and translators
- Model specifications:
  - − Choose **model specifications**
  - \+ Try different models
  - \+ Try **unigram** and **bigram**
  - \+ Try different # words / datapoint
    **100 words** works best

## Data Collection

Six source languages:

**English, Spanish, French, Portuguese, Russian, Korean**

Dataset Specifications:

- **100** words per data point
- **6,000** data points per language

Datapoint Example:

**Text:**

['They were in the garden. Through the back window I could see them before they could see me. Pablo was talking with a glass of white wine in his hand, while his two children looked at him from across the table. His daughter had her elbows on the white tablecloth']

**Label:** 'Spanish'

## Approaches

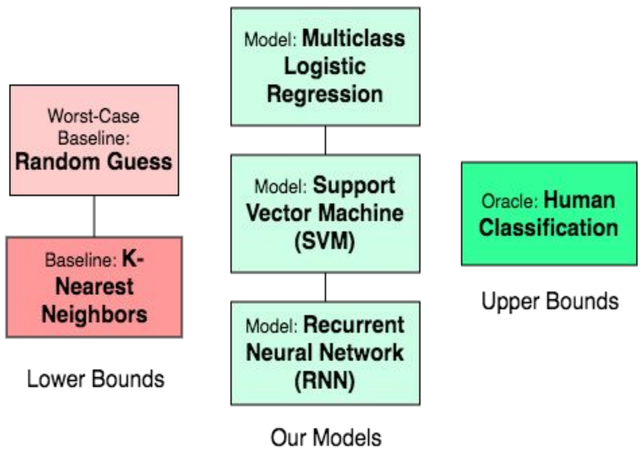


Figure 2. Approaches Summary

## Approaches (cont.)

Upper and Lower Bounds:

- Baselines:
  **K-Nearest Neighbors**: **40.5%** accuracy
  **Random Prediction**: **16.9%** accuracy
- Oracle:
  **Human Classification**: **90.7%** accuracy

Machine Learning Models:

- **Multiclass Logistic Regression**
- **Support Vector Machine** (SVM)
- **Recurrent Neural Network** (RNN):

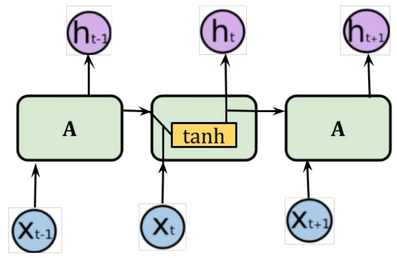


Figure 3. A RNN with a single-layer repeating module

## Results

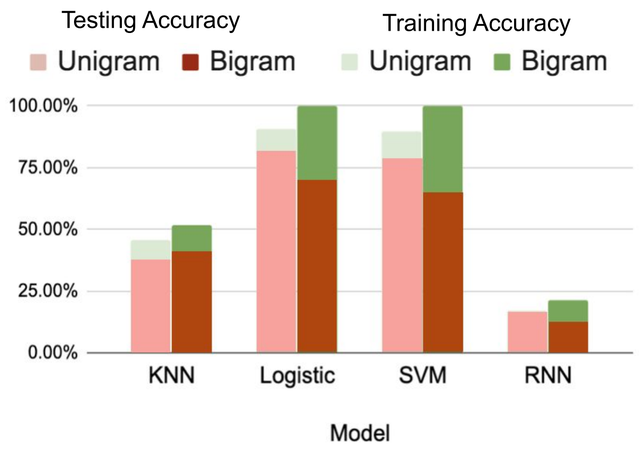


Figure 4. Training and Testing Accuracies for Models

## Results (cont.)

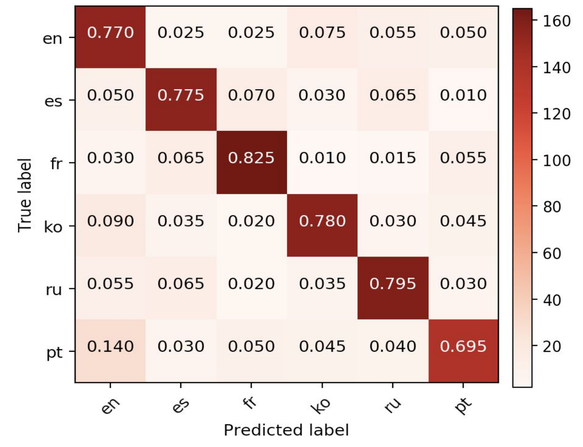


Figure 5. Logistic Regression with Unigram Confusion Matrix

## Analysis

- **Unigram** accuracy higher than bigram
  - Due to **overfitting**
- **Portuguese** most commonly misclassified as **English**
  - **Romance languages** commonly misclassified as each other
- **Logistic Regression** has highest accuracy (marginally over SVM)
  - Logistic loss sensitive to "outliers"
- RNN has poor accuracy:
  - **vanishing gradient**
  - Future Work: solve using **long short term memory network (LSTM)**

### Sources

Baroni, Marco, and Silvia Bernardini. "A new approach to the study of translationese: Machine-learning the difference between original and translated text." *Literary and Linguistic Computing* 21.3 (2005): 259-274.

Lynch, Gerard, and Carl Vogel. "Towards the Automatic Detection of the Source Language of a Literary Translation." *Proceedings of COLING 2012: Posters*. 2012.

Kurokawa, David, Cyril Goutte, and Pierre Isabelle. "Automatic detection of translated text and its impact on machine translation." *Proceedings of MT-Summit XII* (2009): 81-88.