# Portuguese Native Language Identification

**Shervin Malmasi[1], Iria del Rio[2], Marcos Zampieri[3]**

[1]Harvard University, USA
[1]Amazon Inc., USA
[2]University of Lisbon, Portugal
[3]University of Wolverhampton, UK

# Outline

# Introduction

- **Native Language Identification (NLI)** is the task of automatically identifying the native language (L1) of an individual based on their foreign or learned language production (L2).
- Based on L1 interference or language transfer.
- Applying machine learning classifiers to (annotated) learner corpora/L2 data.
- Most work has been done on English. A few papers on other languages: Arabic, Chinese, Finnish, Norwegian.
- No papers published on Portuguese thus far.

# Why?

- Engineering and NLP:
  - Author profiling.
  - Forensics.
  - Tailoring NLP tools to non-native speakers.
- Language learning:
  - Studies in second language acquisition.
  - Educational NLP applications.

# Language Transfer

- Named entities play an important role. Annotation and other strategies can help reducing them for linguistic analysis.
- NLI can be used to study L1 interference (Bykh and Meurers, 2014).
- Positive vs. negative language transfer.
- Examples included in the top 10 most informative features (unigrams and bigrams) for English L2 according to Gebre et al. (2013).
  - Arabic L1: advertisment, statment
  - French L1: exemple, developped
  - Italian L1: infact
  - Spanish L1: necesary, an specific, diferent

# Related Work

- Several studies on English and a few other languages (Jarvis, 2013; Bykh and Meurers, 2014; Malmasi and Dras, 2016).
- Two shared tasks have been organized at BEA:
  - **NLI 2013**[1] (Tetreault et al. 2013): TOEFL 11 dataset (Blanchard, 2013). It contains 11,000 TOEFL essays written by speakers of each of the following L1s: Arabic, Chinese, French, German, Hindi, Italian, Japanese, Korean, Spanish, Telugu, and Turkish.
  - **NLI 2017**[2] (Malmasi et al., 2017): Same 11 L1s as the 2013 edition. 11,000 essays written by speakers of each L1, orthographic transcriptions of 45-second English spoken responses, and iVectors (1,000 instances for each of the eleven native languages).

[1] https://sites.google.com/site/nlisharedtask2013/home
[2] https://sites.google.com/site/nlisharedtask/home

# Dataset

- In previous work we compiled and annotated a dataset called NLI-PT (del Río et al., 2018).
- Written production of learners of European Portuguese.
- Here we use a sub-set of the NLI-PT dataset with texts for five L1 groups: Chinese, English, German, Italian, and Spanish
- L1s with the greatest number of texts in NLI-PT
- Sub-set balanced in terms of proficiency level by L1, texts and tokens

# Dataset

| L1 | Texts | Tokens | Types | TTR |
|---|---|---|---|---|
| Chinese | 215 | 50,750 | 6,238 | 0.12 |
| English | 215 | 49,169 | 6,480 | 0.13 |
| German | 215 | 52,131 | 6,690 | 0.13 |
| Italian | 215 | 51,171 | 6,814 | 0.13 |
| Spanish | 215 | 47,935 | 6,375 | 0.13 |
| Total | 1,075 | 251,156 | 32,597 | 0.13 |

Table: Distribution of the five L1s in the NLI-PT datasets in terms of texts, tokens, types, and type/token ratio (TTR).

# Annotation

- Texts in NLI-PT are automatically annotated using available NLP tools at two levels: Part of Speech (POS) and syntax
- Simple POS: only type of word (LX tagger)
- Fine-grained POS: type of word+morphological features (Freeling)
- Constituency analysis (LX parser)
- Dependency analysis (DepPattern)

# Methods

- Standard multi-class classification approach. Linear Support Vector Machine and feature vectors are created using relative frequency values, in line with previous NLI research (Malmasi and Dras, 2017).
- A single model is trained on each feature type. We then combine all our features using a mean probability ensemble.
- We report our results as classification accuracy under $k$-fold cross-validation, with $k = 10$.

# Features

- Function words: a list of 220 Portuguese function words.
- POS Tags: arranged in n-gram size 1 to 3.
- CFG production rules.
- Traditional character and word n-grams were not used due to thematic bias.
- Orthographic errors and named entities not taken into account.

# Results

| **Feature Type** | **Accuracy (%)** |
|---|---|
| Random Baseline | 20.0 |
| Function Words | 38.5 |
| POS 1-grams | 46.3 |
| POS 2-grams | 52.8 |
| POS 3-grams | 44.9 |
| CFG Production Rules | 43.3 |
| **Ensemble Combination** | **54.1** |

Table: Classification results under 10 fold cross-validation (accuracy is reported).

# Results: Per Class

| Class | Precision | Recall | F1-Score |
|---|---|---|---|
| CHI | 0.571 | 0.796 | 0.665 |
| ENG | 0.507 | 0.326 | 0.397 |
| GER | 0.542 | 0.547 | 0.545 |
| ITA | 0.549 | 0.577 | 0.562 |
| SPA | 0.510 | 0.460 | 0.484 |
| Average | 0.536 | 0.541 | 0.531 |

Table: Ensemble system per-class results: precision, recall and the F1-score are reported.

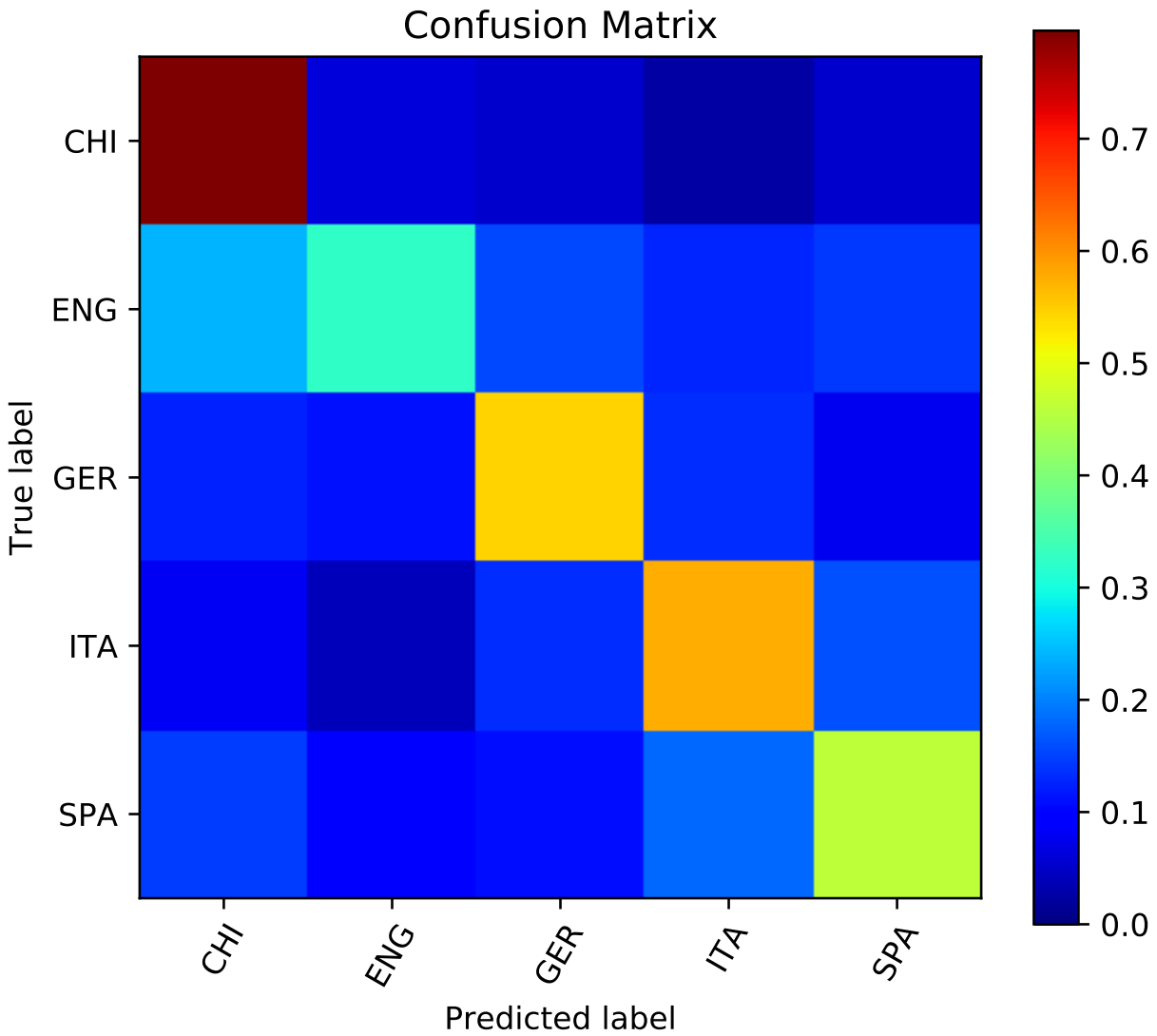


Figure: Confusion matrix for our ensemble system.

# Conclusion

- First study on Portuguese NLI.
- Best results obtained using the ensemble system.
- Best performance on Chinese L1.
- Delexicalized representation.

# Future Work

- Feature analysis.
- Brazilian Portuguese NLI. Anyone?
- Annotation accuracy.
- Proficiency levels.
- Thematic bias.

# Portuguese Native Language Identification

**Shervin Malmasi[1], Iria del Rio[2], Marcos Zampieri[3]**

[1]Harvard University, USA
[1]Amazon Inc., USA
[2]University of Lisbon, Portugal
[3]University of Wolverhampton, UK

# SemEval 2019 Task 6

- If you are interested in offensive language and hate speech please check the **SemEval 2019 Task 6**. Training data available in October.

- **https://competitions.codalab.org/competitions/20011**

- OffensEval: Identifying and Categorizing Offensive Language in Social Media
  - Sub-task A: Offensive language identification;
  - Sub-task B: Automatic categorization of offense types;
  - Sub-task C: Offense target identification.