LT Programming Competition Team: UniMelb LT

Richard Fothergill,[♡] Aidan Nagorcka-Smith[♡] and Li Wang[♠♡]

♠ NICTA Victoria Research Laboratory

♡ Dept of Computer Science and Software Engineering, University of Melbourne

Overview

Competition Description: This competition is formatted as a "shared task" where participants aim to build the best language identification system for multilingual documents.

Competition materials Language-annotated training and development document corpora and an unlabelled test corpus were provided with the task.

Our Best Results: Precision: 0.892; Recall: 0.892; F-score: 0.892

Corpus

Examination of the provided training corpus revealed that:

- Classification instances are mixed-language documents in Wikipedia's Mediawiki syntax.
- Gold standard labels on training data give the set of languages in the document, with no alignment to the text.
- At most two language labels are given per document.
- Documents contain features specific to the domain of Wikipedia text (illustrated in figure below).

Example Document:

```
{{Infobox Settlement
|official_name            = Ferndale, Florida

<!-- Population   ----------------------->
|population_as_of          = 2000
|population_footnotes      =
|population_note           =
|population_total          = 233
|population_density_km2    = 31.5
|population_density_sq_mi  = 83.2

}}

'''Ferndale''' is a Lake County, Florida, United States. The
population was 233 at the 2000 census.

==Geography==
Ferndale is located at  (28.619342, -81.702935).

According to the United States Census Bureau, the CDP has a
total area of 2.8&nbsp;square miles (7.4&nbsp;km²), of which,
2.7&nbsp;square miles (7.1&nbsp;km²) of it is land and
0.1&nbsp;square miles (0.3&nbsp;km²) of it (3.87%) is water.

==Demographics==
As of the White, 0.43% African American, 1.72% from other
races, and 0.43% from two or more races. Hispanic or Latino
of any race were 3.86% of the population.

There were 83 households out of which 34.9% had children
under the age of 18 living with them, ...


{{Info/Localidade dos EUA|<!--Ferndale (Florida)-->
|nome       = Ferndale
|estado     = Flórida
|condado    = Condado de Lake
|população  = 233
|data_pop   = 2000

}}
'''Ferndale''' é uma Estado americano de Condado de Lake.

==Demografia==
Segundo o censo americano de 2000, a sua população era de
233 habitantes<ref>[http://www.census.gov/Press-Release/
www/2001/sumfile1.html U.S. Census Bureau. Census 2000
Summary File 1]</ref>.

==Geografia==
De acordo com o '''United States Census Bureau''' tem uma
área de
7,4 km², dos quais 7,1 km² cobertos por terra e 0,3 km²
cobertos por água.
```

Legend

Template Keys

Section Headings

Machine Learning

Features Extracted:

Byte and unicode N-grams The primary features used were byte bigrams and trigrams.

Wiki headings and template keys As illustrated in the example document to the right these often fall in a small set of words specific to the document language.

Writing System The unicode script database was used to calculate the proportion of each document in each writing sytem.

Multiclass strategies used:

Stratification The full set of languages for a document combined into a single label.

Binarization A single classifier trained for each language.

Native multiclass Many algorithms natively supported multiclass.

Machine learning algorithms used:

kNN/kNP Nearest neighbour and prototype; using cosine similarity and skew divergence; calculating one to five nearest feature neighbours.

SVM Support vector machines with linear and rbf kernels.

NB Multinominal Naïve Bayes.

System: Line by Line

Pre-processing for each document:

1. Remove all URLs.
2. Remove most Media-wiki markup, including most punctuation.
3. Convert the entire document to lower case

Processing of training documents:

1. Extract byte bigrams from the training documents.
2. Stratify classes into language pairs.
3. Produce a prototype for each language pair.

Classification for testing documents:

1. Extract byte bigrams for each individual line within a document.
2. Find five nearest prototypes for each line based on vector cosine similarity.
3. Assign the majority language label within these 5 prototypes to the line.
4. Allow each line within a document to "vote" towards the document language, weighting each vote using the length of the line in bytes.
5. Take the two languages with the largest number of votes, and assign those two labels to the document.

Results

System	Precision	Recall	F-Score
Linear SVM	0.889	0.890	0.890
Combination	0.794	0.793	0.794
Line by Line	0.892	0.892	0.892

System: Linear SVM

Our simplest submission was an SVM learner:

features byte bigrams

classes Stratified (language pairs)

SVM kernel Linear

System: Combination

For another submission, we generated a meta classifier for a combination of basic classifiers.

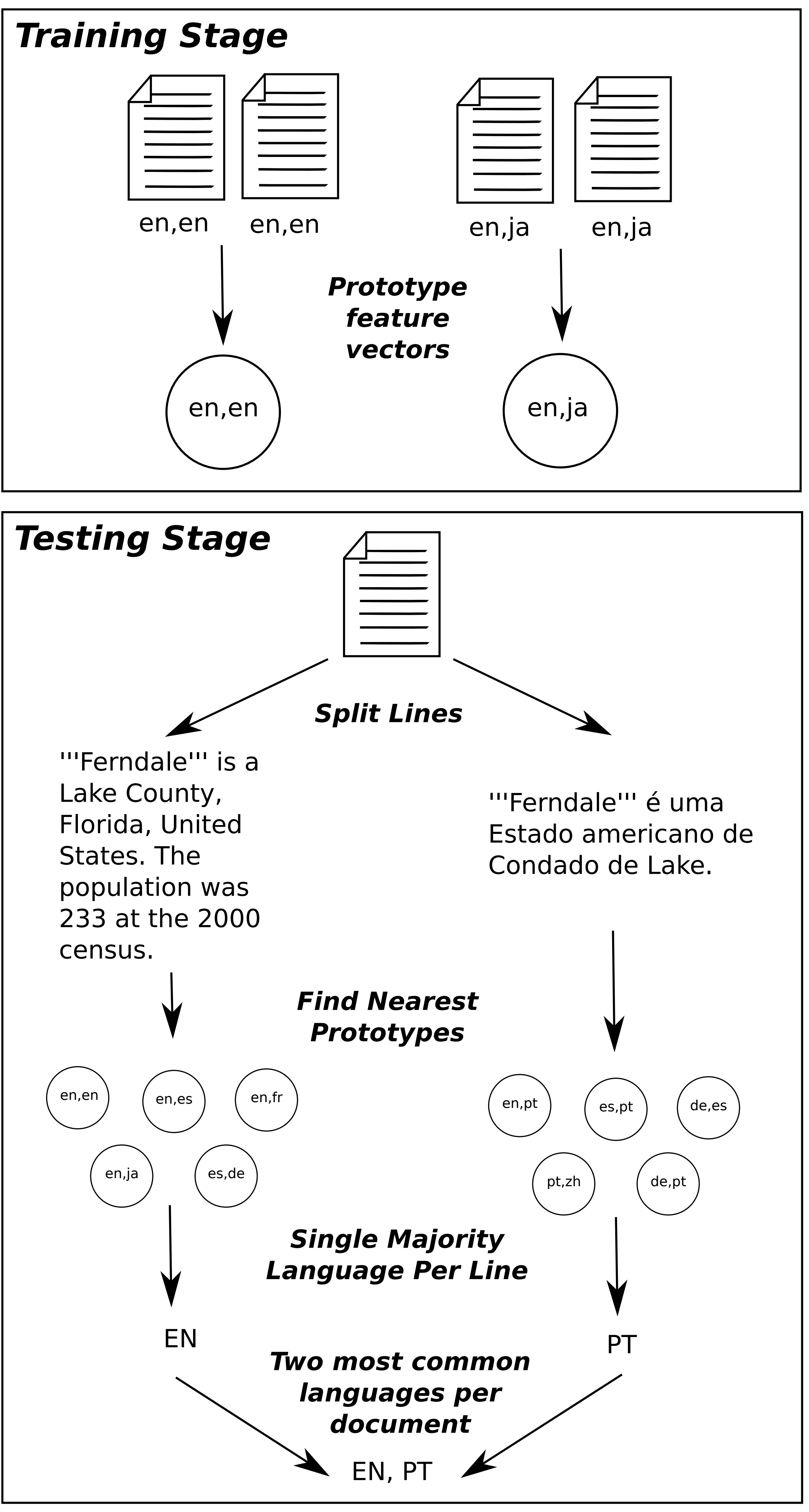
Approach:

- Train different level-0 systems over the training data.
- Combine all the predictions for the development data as level-1 training data.
- Combine all the predictions for the test data as level-1 test data.
- Train a level-1 system over the level-1 training data and classify over the level-1 test data.

In retrospect:

- Naive stacking approach for the convenience of put all the predictions from 3 team members together.
- Preferably do a full and complete stacking based on stratified 10-fold crossvalidation.

Line by line illustration


Training Stage
en,en
en,en
en,ja
en,ja
Prototype feature vectors
en,en
en,ja
Testing Stage
Split Lines
'''Ferndale''' is a Lake County, Florida, United States. The population was 233 at the 2000 census.
'''Ferndale''' é uma Estado americano de Condado de Lake.
Find Nearest Prototypes
en,en
en,es
en,fr
en,ja
es,de
en,pt
es,pt
de,es
pt,zh
de,pt
Single Majority Language Per Line
EN
PT
Two most common languages per document
EN, PT

Acknowledgements

We would like to thank our team mentors, Timothy Baldwin and Marco Lui.