Introduction

- Support Vector Machines are applied to the Native Language Identification task.
- We introduce 2 novel features: Spelling Error Alignments, and Cognate Interference Features, which improve classifier accuracy by 2 and 4%, respectively.
- Our best classifier contains our features, as well as a set of other common features.

Features

- **Word Unigrams:**
  - Function words:
  - A list of 295 common function words such as "for", "the", "a", etc.
  - Content words:
  - We consider all word unigrams encountered in training.
  - Word Bigrams
  - Character Bigrams
- **Binary feature that indicates if bigram occurs in document with greater-than-average frequency**
- **Part-of-Speech Bigrams**
- **Meta-features:**
  - Number of words / sentence
  - Number of characters / document
  - Number of words / document
  - Number of characters / document
  - Average Word Length
- **Syntax Production Rules**
  - Rules of the form A → BC
  - Rules with and without terminals
  - Cognate Interference Rules
  - Spelling Error Features

Features were used to train a multi-class Support Vector Machine with a linear kernel.

Cognate Features

- We hypothesize that spelling is affected by L1.
- If the misspelling is closer to the L1 word than the L2 word is, we activate a cognate feature.
- Cognate features were developed for French, German, Italian, and Spanish.

1. For each misspelled English word \( m \) found in a document, identify the most likely intended word \( e \) using a spell-checking program.
2. For each language L:
   a. Look up the translation \( f \) of the intended word \( e \) in language L.
   b. Compute the orthographic edit distance \( D \) between the words.
   c. If \( D(e,f) < t \) then \( f \) is assumed to be a cognate of \( e \).
   d. If \( f \) is a cognate and \( D(m,f) < D(e,f) \) then we consider it as a clue that \( L = L1 \).

Orthographic distance between misspelling and cognate

- `concentrated` is closer to `konzentrierte` than `concentrated` is.
- **The German feature fires.**

Misspelling | Intended | Cognate
---|---|---
developed | developed | developpé (FRE)
exemple | example | exemple (FRE)
organisation | organization | organisation (GER)
comercial | commercial | comercial (SPA)
sistem | system | sistema (SPA)

Algorithm for determining cognate interference

Examples of cognate interference

Spelling Error Features

- We hypothesize that speakers with the same L1 make similar spelling mistakes.
- Misspellings are aligned with their proposed corrections. An example alignment is shown below.
- Alignments are used as binary features in the SVM: if the error is detected, the feature is 1, otherwise, it is 0.

A misaligned word and its correction

```
<table>
<thead>
<tr>
<th>a</th>
<th>b</th>
<th>s</th>
<th>t</th>
<th>ru</th>
<th>ct</th>
</tr>
</thead>
<tbody>
<tr>
<td>l</td>
<td>l</td>
<td>l</td>
<td>l</td>
<td>l</td>
<td>l</td>
</tr>
</tbody>
</table>
```

Common spellings for each of the languages

ARA | CHI | FRE | GER | HIN | ITA | JPN | KOR | SPA | TEL | TUR
---|---|---|---|---|---|---|---|---|---|---
me -> m | o -> e | n -> nn | ct -> kt | ze -> se | y -> i | r -> l | ur -> u | ss -> s | za -> sa | ci -> si
ne -> n | i -> u | ll -> l | s -> s | e -> ea | ch -> c | l -> r | ed -> d | nm -> m | po -> pu | hy -> y
t -> te | t -> d | xa -> xe | ed -> et | es -> ce | b -> bb | n -> nn | ti -> i | qu -> cu | al -> l | ge -> g

Common spellings for each of the languages

ARA | CHI | FRE | GER | HIN | ITA | JPN | KOR | SPA | TEL | TUR
---|---|---|---|---|---|---|---|---|---|---
me -> m | o -> e | n -> nn | ct -> kt | ze -> se | y -> i | r -> l | ur -> u | ss -> s | za -> sa | ci -> si
ne -> n | i -> u | ll -> l | s -> s | e -> ea | ch -> c | l -> r | ed -> d | nm -> m | po -> pu | hy -> y
t -> te | t -> d | xa -> xe | ed -> et | es -> ce | b -> bb | n -> nn | ti -> i | qu -> cu | al -> l | ge -> g

Accuracy of our best classifiers

- Base
- Corrected Words
- Chains
- Chains + Meta

The results of our best classifiers

- No spelling errors
- No word unigrams
- No cognate features
- No production rules
- No character bigrams
- No word bigrams

Ablation Study

- Best accuracy of 81.7% obtained with all features except POS bigrams.
- Misclassifications often occur between languages with historical connections.
- Both cognates and spelling features improve accuracy of classifier.