Email classification for semi-automated reply generation

I256: Applied Natural Language Processing
Semester project

Hannes Hesse
Background

Youth For Understanding, organizer of student exchange

Sends ~1000 German students abroad each year

Largely run by volunteers, not heavily staffed
Background

www.yfu.de: 10 - 15 inquiries per day
Typical questions

“Please send me application material or brochures”

“I am XY years old. Can I still apply?”

“Is it too late to apply now for year XY?”

“Can I go abroad for six months?”

“We would like to host an exchange student.”

“Where’s my stuff?”
Objective

Classify emails for

- Routing
- Auto-suggest feature in CRM system
- Analysis
Data Preprocessing

1. Convert HTML emails in .mbox format to text files

2. Extract author name and message body, discard headers and HTML

3. Convert special characters, remove capitalization

4. Tokenize words

5. Stemming, using German version of Porter algorithm
Training data

Original idea: Unsupervised clustering for category exploration (n =~ 1800)

Used Orange for data exploration and clustering experimentation

Manual classification of training data (n=350)

Initially: 26 categories, reduced to 10
Distribution of classes
Training data

- bewerbungsunterlagen
- ap_interesse
- ep_teilnahmebestaetigung
- halbes_jahr
- stipendien
- altersgrenzen
- bewerbungsfristen
- ep_unterlagen_statusnachfrage
- intranet_support
- ep_direct_placement

Distribution of classes with Training data graph.
Feature selection

1. Stopword removal

2. Initial selection of 500 most frequent terms

3. Supervised feature reduction to 16 features, using Weka

Both stemming and feature reduction improved accuracy.
Classification
Algorithms and results

Experimented with Naive Bayes, multinomial Naive Bayes and Support Vector Machine.

Naive Bayes yielded best results.

Precision for single classes ranged from 0% - 84%.

Overall correctly classified: ~77%
### Classification accuracy

Confusion matrix

<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
<th>e</th>
<th>f</th>
<th>g</th>
<th>h</th>
<th>i</th>
<th>j</th>
<th>classified as</th>
</tr>
</thead>
<tbody>
<tr>
<td>190</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>6</td>
<td>0</td>
<td>_BEWERBUNGSUNTERLAGEN</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>_altersgrenzen</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>_ap_interesse</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>0</td>
<td>15</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>_bewerbungsfristen</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>_ep_teilnahmebestaetigung</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>10</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>_ep_unterlagen_statusnachfrage</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>25</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>_halbesjahr</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>_intranet-support</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>4</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>3</td>
<td>0</td>
<td>8</td>
<td>0</td>
<td>_stipendien</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>4</td>
<td>_ep_direct-placement</td>
<td></td>
</tr>
</tbody>
</table>

Kappa = .61
Further analysis
What countries are people interested in?

![Bar chart showing website traffic and email occurrences for various countries.](chart.png)
Further analysis

Wording: What do people call the product?

- Auslandsschuljahr
- High School Year
- Programme
- Austauschprogramm
- Auslandsaufenthalt
- Schüleraustausch
- Auslandsjahr
- Austausch
- Austauschjahr

0 8 15 23 30
Further analysis
Recommendations for site usability refinement

People don’t find order form for application material

Perhaps site does not communicate details about programs clearly enough
Results and Outlook

Some emails are easy to handle, others need human attention

For from reliable auto-reply generation

Larger training set needed

Examine correlation between incoming and actual replies

Implement some email routing