Native Language Identification

A challenging task description

Barbora Hladká                Martin Holub

{hladka | holub}@ufal.mff.cuni.cz

Charles University,
Faculty of Mathematics and Physics,
Institute of Formal and Applied Linguistics
This indicates that to be specialized in one specific subject is also a good thing it will make you a professional in a specific subject, but when you come to society they will.

Changing your habits is not an easy thing, but one is urged to do it for a number of reasons. A successful teacher should renew his lectures so.

It is troublesome, as it seems, but it keeps me fresh in information and with a good status among my colleagues.

Also, you might think of changing your view, clothes or even your hair cut.

Lastly, change is a difficult decision in the human, but it is important for many good reasons, and gets back on the human with good benefits.

To be specialized in one specific subject is also a good thing it will make you a professional in a specific subject, but when you come to society they will.

Alternating work and dormancy in your life pace with activities and exercise is of great benefit to the body and the mind.

In contrast, many people believe that changing is really important in people’s lives.

When I change my glasses color, for example, this would be attractive to my students and colleagues and will make me feel better.

The newer method is considered an invention in its own, and it also saves money.

Lastly, change is a difficult decision in the human, but it is important for many good reasons, and gets back on the human with good benefits.

Changing your habits is not an easy thing, but one is urged to do it for a number of reasons. A successful teacher should renew his lectures so.

Also, you might think of changing your view, clothes or even your hair cut.

Lastly, change is a difficult decision in the human, but it is important for many good reasons, and gets back on the human with good benefits.

In contrast, many people believe that changing is really important in people’s lives.

When I change my glasses color, for example, this would be attractive to my students and colleagues and will make me feel better.

The newer method is considered an invention in its own, and it also saves money.

Lastly, change is a difficult decision in the human, but it is important for many good reasons, and gets back on the human with good benefits.
Sample texts

(1) Many people lose their life, their job, or maybe their family. All of them were not learn from other’s fault. Successful person who has many choses that help him in his life. So, what we lose if we learn one more thing in our life.

(2) Last week I read an article in our daily newspaper about who enjoys life more, female or male. I considered about these piece of paper a very long time and came to the conclusion that it is more worth to distinguish between young and old people, than between female and male.
Sample texts

(1) Many people lose their life, their job, or maybe their family. All of them were not learn from other’s fault. Successful person who has many choses that help him in his life. So, what we lose if we learn one more thing in our life.

(2) Last week I read an article in our daily newspaper about who enjoys life more, female or male. I considered about these piece of paper a very long time and came to the conclusion that it is more worth to distinguish between young and old people, than between female and male.

(1) ... ARA
(2) ... GER
Native Language Identification

Task: Predict L1 of English essays’s authors

Data:

• TOEFL11 is a corpus of non-native English writing
  • consists of essays on eight different topics (prompts P1 to P8)
  • written by non-native speakers of three proficiency levels (labels low/medium/high)
  • the essays’ authors speak 11 different native languages that should be predicted
  • contains 1,100 tokenized essays per language with an average of 348 word tokens per essay
  • more info can be found in (Blanchard et al., 2013)
• Additionally, all texts have been preprocessed by the Stanford POS tagger (Toutanova et al., 2003).
## Existing approaches to NLI

### State of the art results

<table>
<thead>
<tr>
<th>System</th>
<th># of feat.</th>
<th>Acc.*</th>
<th>Approach</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gebre et al., 2013</td>
<td>73,626</td>
<td>84.6</td>
<td>tf-idf of unigrams and bigrams of words</td>
</tr>
<tr>
<td>Jarvis, Bestgen, Pepper, 2013</td>
<td>400K</td>
<td>84.5</td>
<td>{1,2,3}-grams of words, lemmas, POS tags, df ≥ 2</td>
</tr>
<tr>
<td>Kríž, Holub, Pecina, 2015</td>
<td>55**</td>
<td>82.4</td>
<td>language models using tokens, characters, POS, suffixes</td>
</tr>
</tbody>
</table>

*Acc. – cross-validation results on train+dtest

**traditional n-grams are hidden in the language models
Analyzing most discriminative n-grams

A sample of extracted word $n$-grams

<table>
<thead>
<tr>
<th>$n$-gram</th>
<th>fr</th>
<th>df</th>
<th>G max</th>
<th>G max$_2$</th>
<th>G max$_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>,</td>
<td>114,358</td>
<td>8,701</td>
<td>-2,857.8</td>
<td>2,328.7</td>
<td>2,239.8</td>
</tr>
<tr>
<td>i</td>
<td>6,011</td>
<td>1,939</td>
<td>992</td>
<td>957.4</td>
<td>735.7</td>
</tr>
<tr>
<td>Japan</td>
<td>356</td>
<td>210</td>
<td>982.1</td>
<td>380.8</td>
<td>379.2</td>
</tr>
<tr>
<td>think</td>
<td>9,883</td>
<td>4,861</td>
<td>-915.1</td>
<td>797.6</td>
<td>-767.9</td>
</tr>
<tr>
<td>think that</td>
<td>3,602</td>
<td>2,322</td>
<td>807.7</td>
<td>748.5</td>
<td>506.2</td>
</tr>
<tr>
<td>I think that</td>
<td>1,963</td>
<td>1,375</td>
<td>796.3</td>
<td>523.6</td>
<td>391.8</td>
</tr>
<tr>
<td>tour</td>
<td>3,810</td>
<td>694</td>
<td>-725.3</td>
<td>-644.7</td>
<td>-625.3</td>
</tr>
<tr>
<td>Indeed</td>
<td>282</td>
<td>222</td>
<td>694.2</td>
<td>288.5</td>
<td>282.4</td>
</tr>
<tr>
<td>in twenty</td>
<td>1,214</td>
<td>703</td>
<td>73.1</td>
<td>-62.8</td>
<td>NA</td>
</tr>
</tbody>
</table>

fr . . . absolute n-gram frequency in the corpus

df . . . document frequency (number of examples containing given n-gram)

G stands for $G$-test statistic.
Language modeling approach to feature extraction

- We built 11 special language models of English \((M_i)\), each based on the texts with the same L1 language available in the training data.
- Then we compare \(M_i\) to a general language model of English \((M_G)\).
- The cross-entropy of text \(t\) with empirical n-gram distribution \(p\) given a language model \(M\) with distribution \(q\) is
  \[
  H(t, M) = - \sum_x p(x) \log q(x).
  \]
- Normalized cross-entropy scores – used as features
  \[
  D_G(t, M_i) = H(t, M_i) - H(t, M_G) = - \sum_x p(x) \log \frac{q_i(x)}{q_G(x)},
  \]
  where \(M_i\) are the special language models with distributions \(q_i\), and \(M_G\) is the general language model with the distribution \(q_G\).
## Our best model – error analysis

### Aggregated confusion matrix

Sum of 10 confusion matrices obtained in 10-fold cross validation process

<table>
<thead>
<tr>
<th></th>
<th>ARA</th>
<th>DEU</th>
<th>FRA</th>
<th>HIN</th>
<th>ITA</th>
<th>JPN</th>
<th>KOR</th>
<th>SPA</th>
<th>TEL</th>
<th>TUR</th>
<th>ZHO</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARA</td>
<td>738</td>
<td>8</td>
<td>30</td>
<td>24</td>
<td>5</td>
<td>16</td>
<td>9</td>
<td>30</td>
<td>10</td>
<td>32</td>
<td>15</td>
</tr>
<tr>
<td>DEU</td>
<td>7</td>
<td>836</td>
<td>13</td>
<td>4</td>
<td>20</td>
<td>4</td>
<td>6</td>
<td>20</td>
<td>1</td>
<td>16</td>
<td>1</td>
</tr>
<tr>
<td>FRA</td>
<td>31</td>
<td>6</td>
<td>767</td>
<td>0</td>
<td>35</td>
<td>6</td>
<td>3</td>
<td>39</td>
<td>1</td>
<td>15</td>
<td>1</td>
</tr>
<tr>
<td>HIN</td>
<td>19</td>
<td>1</td>
<td>1</td>
<td>684</td>
<td>0</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>194</td>
<td>9</td>
<td>4</td>
</tr>
<tr>
<td>ITA</td>
<td>6</td>
<td>8</td>
<td>20</td>
<td>1</td>
<td>761</td>
<td>0</td>
<td>1</td>
<td>58</td>
<td>0</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>JPN</td>
<td>7</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>709</td>
<td>109</td>
<td>6</td>
<td>1</td>
<td>8</td>
<td>22</td>
</tr>
<tr>
<td>KOR</td>
<td>14</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>120</td>
<td>676</td>
<td>6</td>
<td>2</td>
<td>12</td>
<td>50</td>
</tr>
<tr>
<td>SPA</td>
<td>30</td>
<td>21</td>
<td>45</td>
<td>6</td>
<td>69</td>
<td>3</td>
<td>4</td>
<td>720</td>
<td>1</td>
<td>17</td>
<td>7</td>
</tr>
<tr>
<td>TEL</td>
<td>4</td>
<td>2</td>
<td>0</td>
<td>157</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>685</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>TUR</td>
<td>32</td>
<td>15</td>
<td>11</td>
<td>16</td>
<td>6</td>
<td>11</td>
<td>23</td>
<td>11</td>
<td>5</td>
<td>778</td>
<td>16</td>
</tr>
<tr>
<td>ZHO</td>
<td>12</td>
<td>0</td>
<td>7</td>
<td>2</td>
<td>1</td>
<td>28</td>
<td>66</td>
<td>5</td>
<td>0</td>
<td>8</td>
<td>781</td>
</tr>
</tbody>
</table>
## Initial experiments with NN

<table>
<thead>
<tr>
<th>Model</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Paragraph vector model – Distributed Bag-Of-Words</td>
<td>(Le &amp; Mikolov, 2014)</td>
</tr>
<tr>
<td>Paragraph vector model – Distributed Memory</td>
<td>(Le &amp; Mikolov, 2014)</td>
</tr>
<tr>
<td>Word-to-Vec Inversion</td>
<td>(Taddy, 2015)</td>
</tr>
<tr>
<td>Recurrent Neural Network</td>
<td>(Cho et al., 2014)</td>
</tr>
<tr>
<td>Convolutional Neural Network</td>
<td>(Kim, 2014)</td>
</tr>
</tbody>
</table>
Goals

Build and tune a deep neural model to beat the state of the art using

- recurrent neural networks
- convolutional neural networks
- ...

Then the NN model(s) should be compared to the traditional SVM classifiers:

- what is the difference?
- is their performance complementary?


