600.103 - Homework for Week 11/06/03 - Due 11/13/03
Natural Language Processing and Machine Learning (Yarowsky)

On a separate sheet of paper, identify two problems in natural language processing that are of the general type discussed in class today (i.e. classification problems with a relatively small set of fixed classes per word). Please don’t repeat the exact problems described today. Creativity and unusual ideas are strongly encouraged. For each problem:

1. Very briefly describe the problem (a few words will do).
   
   For example, capitalization restoration from single-case text.

2. What are the classes you wish to partition your examples into (i.e. the list of class labels)?
   
   For example, Turkey (country) or turkey (bird)

3. Identify a set of features that will be informative/predictive in this classification.
   
   For example:
   
   Word_to_the_right = SANDWICH ⇒ turkey
   Word_to_the_left = FROM ⇒ Turkey
   Word_found_in_same_sentence = STUFFING ⇒ turkey
   Word_found_in_same_sentence = ANKARA ⇒ Turkey

4. What are the potential sources of training data and how can the class labels be automatically acquired?
   
   For example, take the occurrences of Turkey/turkey in a text, e.g.:
   
   ..Greece, Turkey and Iran collaborated..
   ..chicken, turkey and ham for dinner..

   label each example with the observed capitalization pattern, and then convert all surrounding words to upper case for use as training features. For example:

<table>
<thead>
<tr>
<th>Label</th>
<th>Training examples (words in context)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Turkey</td>
<td>..GREECE, TURKEY AND IRAN COLLABORATED..</td>
</tr>
<tr>
<td>Turkey</td>
<td>..EASTERN TURKEY AFTER THE EARLIER..</td>
</tr>
<tr>
<td>turkey</td>
<td>..CHICKEN, TURKEY AND HAM FOR DINNER..</td>
</tr>
<tr>
<td>turkey</td>
<td>..GRADE-A TURKEY PRICES ROSE ONLY..</td>
</tr>
</tbody>
</table>

   In this case, no hand tagging of the examples was required. The labels for the training data (Turkey or turkey) were extracted automatically from the original text.