Abstract

This report contains data and answers to the questions on named entity tagging assignment.

Introduction

Given the training pair of text and its HTML tagging based on 7 types of entities I have wrote the netagger.pl that tags an arbitrary text from STDIN into the color-tagged HTML file. I have also wrote the evaluation software to compare two color-tagged HTML files and report the statistics for the three types of matches eval.pl. Below are the answers for the questions 1–4.

Answers to the questions

Question 1 (Results)

Training data

Results of running the tagger on the development file ne.dev.txt:

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Typed recall</td>
<td>0.969387755102041</td>
</tr>
<tr>
<td>Typed precision</td>
<td>0.969387755102041</td>
</tr>
<tr>
<td>Labeled recall</td>
<td>0.969387755102041</td>
</tr>
<tr>
<td>Labeled precision</td>
<td>0.969387755102041</td>
</tr>
<tr>
<td>Partial recall</td>
<td>1</td>
</tr>
<tr>
<td>Partial precision</td>
<td>1</td>
</tr>
</tbody>
</table>

As required the partial recall of 1 has been achieved for the development data. Moreover, partial precision is 1. More detailed data for each type of the tag is shown below. Here tmatch stands for number of typed matches, lmatch is the number of labeled matches, and pmatch is the number of partial labeled matches for each particular type.
Person:
tmatch: 7 lmatch: 7 pmatch: 7
test_person: 7 guess_person: 7

Organization:
tmatch: 58 lmatch: 58 pmatch: 61
test_organization: 61 guess_organization: 61

Location:
tmatch: 6 lmatch: 6 pmatch: 6
test_location: 6 guess_location: 6

Date:
tmatch: 15 lmatch: 15 pmatch: 15
test_date: 15 guess_date: 15

Time:
tmatch: 1 lmatch: 1 pmatch: 1
test_time: 1 guess_time: 1

Money:
tmatch: 6 lmatch: 6 pmatch: 6
test_money: 6 guess_money: 6

Percent:
tmatch: 2 lmatch: 2 pmatch: 2
test_percent: 2 guess_percent: 2

Total:
tmatch_total: 95 lmatch_total: 95 pmatch_total: 98

Test data
Results of running the tagger on the test file ne_test.txt:

<table>
<thead>
<tr>
<th>Type</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Typed recall</td>
<td>0.315068493150685</td>
</tr>
<tr>
<td>Typed precision</td>
<td>0.741935483870968</td>
</tr>
<tr>
<td>Labeled recall</td>
<td>0.315068493150685</td>
</tr>
<tr>
<td>Labeled precision</td>
<td>0.741935483870968</td>
</tr>
<tr>
<td>Partial recall</td>
<td>0.397260273972603</td>
</tr>
<tr>
<td>Partial precision</td>
<td>0.935483870967742</td>
</tr>
</tbody>
</table>

More detailed data for each type of the tag is shown below. Here tmatch stands for number of typed matches, lmatch is the number of labeled matches, and pmatch is the number of partial labeled matches for each particular type.

Person:
tmatch: 0 lmatch: 0 pmatch: 0
test_person: 10 guess_person: 0

**Organization:**
tmatch: 5 lmatch: 5 pmatch: 5
test_organization: 26 guess_organization: 5

**Location:**
tmatch: 0 lmatch: 0 pmatch: 0
test_location: 6 guess_location: 0

**Date:**
tmatch: 5 lmatch: 5 pmatch: 9
test_date: 16 guess_date: 11

**Time:**
tmatch: 2 lmatch: 2 pmatch: 2
test_time: 2 guess_time: 2

**Money:**
tmatch: 11 lmatch: 11 pmatch: 13
test_money: 13 guess_money: 13

**Percent:**
tmatch: 0 lmatch: 0 pmatch: 0
test_percent: 0 guess_percent: 0

**Total:**
tmatch_total: 23 lmatch_total: 23 pmatch_total: 29

**Question 2: Summary of the mistakes**

As one can see from the previos subsection, the worst performance on the test data was report for ENAMEX entities. Basically recall is 0 for person and location types. This is easy to explain by the fact that the tagging of these entites was based purely on the lexicon build from `ne_dev.html`.

A few organizations got picked up since they overlapped with the training text.

The recall and precision rates on date, time, and money was relatively high. Among these date has proven to be the most challenging due to less structure and greater variety of possible date entities. The false positive like “BC-TV-SAT-DISHES-NYTSF” is hard to fix, since this expression is not a part of English but rather a code structure.

**Question 3: How to improve the NE tagger?**

For this version of the tagger I primarily focused on recognizing more structured entities: TIMEX and NUMEX. Even current version shows high recall and precision rates for these labeled classes. I think even higher rates can still be achieved using
only regular expressions on surface lexical features. The limitation of this approach, however, has been obvious in the example “BC-TV-SAT-DISHES-NYTSF” since simple lexical features indicate this as TIMEX. A consideration of an alternative identifier that would relate semantics (or frequencies of joint appearances) of SAT and DISHES would be helpful in this case.

There is a lot to be improved for identification of ENAMEX entities. When relying on regular expressions techniques, an extensive lexicon is required. One may argue that I could use such features as capitalization, but while capitalization is somewhat indicative of the ENAMEX expressions, it is not a reliable feature for distinction between types (person, organization, location). Indeed, my experiments with capitalization shown increased total recall to 0.520 but decreased precision down to 0.703.

Instead, or in addition to relying on these surface lexical features, a more fruitful approach could include using syntactic and morphological features, as well as statistical correlations (N-grams).

**Question 4: FSA diagram (please refer to the hard copy)**