Number Sense Disambiguation

Stuart Moore

Supervised by:

Anna Korhonen (Computer Lab)
Sabine Buchholz (Toshiba Research Europe)
Introduction

- Number Sense Disambiguation
- Related Work
- My System
- Error Analysis
- Improvements Post Error Analysis
Number Sense Disambiguation

● Classifying numbers into different senses
  – e.g. year, time, telephone number...

● Similar to Word Sense Disambiguation
  – but not quite the same
Aim

- To successfully classify numbers into sense categories
- To use a semi-supervised method
  - Avoids the need for a large, human annotated training set
  - Allows economical adaptation to different languages and domains
<table>
<thead>
<tr>
<th>Label</th>
<th>Description</th>
<th>Examples</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>NUM</td>
<td>Number (Cardinal)</td>
<td>12, 45, 1/2, 0.6</td>
<td>21253 (56.53%)</td>
</tr>
<tr>
<td>NYER</td>
<td>Year(s)</td>
<td>1998, 80s, 1900s, 2003</td>
<td>7659 (20.37%)</td>
</tr>
<tr>
<td>NORD</td>
<td>Number (Ordinal)</td>
<td>May 7, 3rd, Bill Gates III</td>
<td>3264 (8.68%)</td>
</tr>
<tr>
<td>MONEY</td>
<td>Money (US or other)</td>
<td>$3.45, HK$300, Y20,000, $200K</td>
<td>2909 (7.74%)</td>
</tr>
<tr>
<td>NIDE</td>
<td>Identifier</td>
<td>747, 386, I5, pc110, 3A</td>
<td>1027 (2.73%)</td>
</tr>
<tr>
<td>NTEL</td>
<td>Telephone number (or part of)</td>
<td>212 555-4523</td>
<td>507 (1.35%)</td>
</tr>
<tr>
<td>NTIME</td>
<td>A (compound) time</td>
<td>3:20, 11:45</td>
<td>440 (1.17%)</td>
</tr>
<tr>
<td>NDATE</td>
<td>A (compound) date</td>
<td>2/2/99, 14/03/87 (or US) 03/14/87</td>
<td>307 (0.82%)</td>
</tr>
<tr>
<td>NDIG</td>
<td>Number as digits</td>
<td>Room 101</td>
<td>74 (0.20%)</td>
</tr>
<tr>
<td>NADDR</td>
<td>Number as street address</td>
<td>45 North Street, 5000 Pennsylvania Ave</td>
<td>69 (0.18%)</td>
</tr>
<tr>
<td>NZIP</td>
<td>Zip code or PO box</td>
<td>91020</td>
<td>66 (0.18%)</td>
</tr>
<tr>
<td>PRCT</td>
<td>Percentage</td>
<td>75%, 3.4%</td>
<td>21 (0.06%)</td>
</tr>
</tbody>
</table>

(Counts are from the training data of the North American News Text Corpus)
Overview of my system

- Provide a small amount of labelled training data, together with a large amount of unlabelled training data.
- Iterate, so the output from the first model becomes the training data for the second etc.
- Follows work of Yarowsky, with modifications:
  - I have added extra features to describe numbers and punctuation
  - He uses a POS tagger, I have not (so far).
  - My system can use a wider range of rules than his.
  - My first iteration behaves differently.
Features

• The context of each number is examined for a list of features.
• Local context: ± 5 tokens from the number
  – Punctuation, words, word stems, number features
  – Specific location (e.g. token following number)
• Wider context: ± 15 tokens from the number
  – Words and Word stems only
  – Bag of words (anywhere within the window)
Rules

• Each rule is conditional on the presence of one or two features
  – Consider all possible combinations of features that occur together at least five times in the training corpus.

• Based on Yarowsky's rules, but more powerful
  – He had 'Bag of word' rules, and rules combining two words at positions (-2, -1), (-1, 1), (1, 2)
  – He did not have any specific numeric or punctuation features.
Ranking Rules

- Follows Yarowsky (1995)
- For each rule, count the number of examples for each number sense
- Calculate Log Likelihood:
  \[ \text{LogLike}=\log\left( \frac{\text{Count Positive Examples}}{\text{Count Negative Examples} + \alpha} \right) \]
  - \( \alpha \) is a parameter that can be varied to change the effect of negative examples on the model
- Rank rules according to log likelihood
- When classifying, use the first rule that matches the target sentence
Performance as a semi-supervised system

- After experimenting with a few different strategies for the first iteration, the following was found to perform best:
  - Rank all rules based on their scores from the seed examples
  - For each number type, take the three highest scoring rules (more if several had an equal score)
  - Apply these rules to the unlabelled data.

    - If a number is matched by rules from more than one number type, do not classify it
How many seed examples are needed?

- Seed examples were randomly picked from the training data
- Equal numbers of seed examples for each number type
- Definite improvement seen for going up to 40 seed examples
- Limited improvement after this point
Second Iteration Performance

Peak – 84%
(Loglike cutoff = 5.2)

Training Data
Second Iteration Performance

![Graph showing the relationship between Log Likelihood Cutoff and performance. The peak performance is at 75%, with a cutoff of 5.0. The graph also shows the performance drop-off at higher cutoff values.]
Error Analysis
First Iteration

- We analysed 100 misclassifications from the first iteration
- Most common: Numbers incorrectly being labelled as ordinals because they were preceded by a full stop.
  - The Federal Reserve might have to be more aggressive in raising interest rates at its meeting on Nov. 15, maybe by more than a half of a percentage point
Error Analysis
First Iteration

• This lead to a rule that a two digit number following a full stop was labelled as a NORD
  – This category is for ordinal numbers, used for 'day of month' in the Sproat labelling scheme

• However, this leads to a misclassification of 35% of examples, such as
  – ... placing it at No. \textbf{30} on Forbes' recent ranking ...
Error Analysis
First Iteration

● Twenty of these examples appear to be table data in the original newspaper articles, where the data made little sense without the tabular formatting.

  – THE BEAUTY OF THE LILIES, by John Updike. .. 10 ... - ... 2 ( Knopf, $25.95. ) Over four generations, a family shifts its allegiance from religious to cinematic illusions.
Error Analysis
First Iteration

• The next most common error (17%) was from numbers followed by a colon being identified as times.
  - Its current prediction for **1996**: Clinton wins with about 53 percent of the major-party popular vote.

• All seed example times were of the form 20:00, and so the system correctly identified that the colon was an indicator of the NTIME class.
  - However it hadn't identified that the colon had to be wholly within the number.
Error Analysis
First Iteration

• At present there are no suitable features to capture the difference between this and a colon appearing at the end of a token
Error Analysis
First Iteration

• The next most common error (10%) was due to an unclassified number being adjacent to another four digit number. This rule originated from examples such as "November 1, 1994", but causes misclassifications such as
  – United States v. Cochran, 883 F. 2d 1012, 1017-1018 (CA11 1989);

• All occurrences of this kind of error were either from references, as above, or from tables of financial data.
Error Analysis
First iteration

• 8 examples were found where the corpus labelling was in error (and the system was correct)
  – since the MRTA assault on Dec. 17, even these visits have been suspended. *Example is labelled as NUM in the corpus*

• The remainder of the examples performed poorly due to a few words that happened to coincide in the training examples.
  – Often article metadata led to confusion
Error Analysis
First Iteration

- Many of the ordinals occurred near a month (or an abbreviation of one) – however, because there are so many months, none of them occurred often enough to make it into the final ruleset.
  - This is fixed by adding a 'month' class – discussed later.
Error Analysis
Second Iteration

- 100 examples were examined:
- The presence of various items of punctuation (especially comma and colon) as part of the number leads to 24 labelling errors.
  - Dallas 28, Arizona 21: When Troy Aikman goes out with a concussion, backup Rodney Peete comes off the bench
- Most of these are again because the system is unable to express the difference between punctuation within and following a number.
Error Analysis
Second Iteration

• The next most common error (13) is from rules relating to double-dashes – punctuation that mainly occurs within article metadata.

  – ( Lloyd DeGrane/New York Times Photo. )
    ( NYT13 ) CHICAGO. -- Jan. 1, 1995 --
    CALIF-LIBRARY-HOMELESS-1, 1-1 -- The San Francisco Main Library is vacating its Beaux Arts building
Error Analysis
Second Iteration

• Next (12) is from full stops in the token preceding the number, as observed in the first iteration.
  – This is often combined with other features, e.g. the number having 2 digits; the system has picked up that a . before is not sufficient information.
Error Analysis
Second Iteration

- The next most common (10 occurrences) is that the word 'in' preceding the target number is assumed to indicate a year.
  - Dorothy debuts in 101st
- The remaining errors were difficult to assign to broader categories.
Changes post Error Analysis

- Windowed features that occur for more than 20% of Numbers were removed
  - Gets rid of 'a', 'the'
  - Only windowed features; 'the' occurring at a specific position is still counted.

- Improved way of expressing punctuation within numbers
  - Can now express 'digit digit colon digit digit' for Time
Changes post Error Analysis

• Introduce classes for Months, and Days of the Week.
  – Include abbreviations
  – Motivation: Clearly useful for date based procedures
  – Closed classes, domain independent and easy to generate for other languages.

• Possibly do the same for ordinal suffixes.
Changes post Error Analysis

Performance on Training Data

Peak – 87%
Cutoff - 6.1

LogLike Cutoff

% Correct
% Correct (treating unlabelled as NUM)
Changes Post Error Analysis

• Use more specific rules
  – If Rules X and Y have the same score, which has come from the same labelled examples
  – And all numbers (including unlabelled ones) matching rule X also match rule Y
  – Then remove rule Y

• Reasoning: X better describes the situation than Y does

• Danger: Decision lists rely on these less useful rules
Next Steps

- Error analysis of the new system, and make any obvious improvements
- Define new set of senses for numbers
- Produce a new corpus.
  - Newswire Text
  - Email
- Investigate getting people to highlight the useful context