Interactive Information Extraction

Trausti Kristjansson, Aron Culotta, Paul Viola, Andrew McCallum

Introduction

In USA, 70 millions workers complete forms on a regular basis.

The goal of this work is to reduce the burden on the user to the largest extent possible, while ensuring the integrity of the data.

Main Points

• Synergy of User Interface and Information Extraction Algorithm
• CRFs for information extraction
• Correction Propagation in CRFs
• Confidence Estimation in CRFs
• Expected Number of User Actions

Add Contacts to Address Book

• Email
• Web
• Text document
• Word document
• Excel

Demo: Contact Assistant

Data Integrity – Fast Verification

Input is automatically parsed and assigned to fields.

Color coded correspondence, user can quickly spot errors.
Correction Propagation
• Show live demo

Interactive Information Extraction
• UI shows automatic field assignment results and allows for fast verification and fast correction
• IE algorithm takes corrections into account and propagates correction to other fields
• IE algorithm calculates confidence scores
• UI uses confidence scores to alert user to possible errors

Constrained Conditional Random Fields and Confidence Estimation

Classes – Database Fields
• Classes
  – First Name
  – Last Name
  – Title
  – Suffix
  – Company Name
  – Phone - Business
  – Phone - Home
  – Phone – Mobile
  – FAX
  – Address Line
  – City
  – State
  – Postal Code
  – Country
  – Email address
  – Webpage URL

Classes – Database Fields
• Classes
  – First Name
  – Last Name
  – Title
  – Suffix
  – Company Name
  – Phone - Business
  – Phone - Home
  – Phone – Mobile
  – FAX
  – Address Line
  – City
  – State
  – Postal Code
  – Country
  – Email address
  – Webpage URL
Classes – Database Fields

• Classes
  - First Name
  - Last Name
  - Title
  - Suffix
  - Company Name
  - Phone – Business
  - Phone – Home
  - Phone – Mobile
  - FAX
  - Address Line
  - City
  - State
  - Postal Code
  - Country
  - Email address
  - Webpage URL

Conditional Random Fields

• Conditional Random Fields are globally normalized probability models, where hidden variables are conditioned on observed variables.

  Do not model the distribution over the observed variables, as generative models do.

  Advantage over generative models (e.g. HMMs) is that independence of observations not necessary.

Token Features \( f_k(y,t) \)

• Features
  - Capitalized
  - All Caps
  - In First Name Lexicon
  - In Last Name Lexicon
  - 1st Word on line
  - 2nd Word on line
  - 3rd Word on line
  - Previous Token in First Name lexicon
  - Contains Digits
  - Contains S Digits
  - Contains Hyphen
  - Enclosed in Brackets

  ... and 20000 more
**Conditional Random Fields**

\[ p(x|y) = \frac{1}{Z_0} \exp \left( \sum_t \sum_y \gamma_t g(x,y) + \sum_t \sum_y \lambda_t f(x,y,t) \right) \]

Normalizing factor, i.e., sum over all state sequences for given observation.

**Finding the best state assignment**

Viterbi used to find best sequence

- Viterbi algorithm may return the sequence of states shown below

**Correction Propagation**

User Correction

- User corrects a field, e.g., dragging Stanley to the Last Name field
Remove Paths

- User Corrects a field, e.g. dragging Stanley to the Last Name field.

Constrained Viterbi

- Viterbi algorithm is constrained to pass through the designated state.

Indicate Low Confident

Confidence Estimation

- Confidence in a classification
- Constrained Forward algorithm used to calculate sum of subset of paths that “agree” and “disagree” with a classification

\[
CE = \frac{P(\text{Classification})}{P(\text{Any classification})} = \frac{\text{Sum of all paths that agree with classification}}{\text{Sum of all paths}}
\]

Sum of “agreeing” states sequences

- Paths that “agree” with classification

Sum of all state sequences

- All paths
Evaluation

Standard Metrics

- Standard information retrieval metrics:

<table>
<thead>
<tr>
<th></th>
<th>Token Acc.</th>
<th>F1</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>CRF</td>
<td>89.73</td>
<td>87.23</td>
<td>88.24</td>
<td>86.24</td>
</tr>
<tr>
<td>MaxEnt</td>
<td>88.43</td>
<td>84.84</td>
<td>85.05</td>
<td>84.95</td>
</tr>
</tbody>
</table>

- These metrics don’t relate well to the stated goals, e.g. how much does the system speed up data acquisition.

Expected Number of User Actions

- UI designers often use the “Number of Clicks” as an objective metric.
- We would like a similar metric for measuring the effectiveness of Correction Propagation.
- We can calculate the Expected Number of User Actions (ENUA) from statistics of the number of erroneous fields in each record processed by the system.

\[
ENUA_{\text{manual}} = \frac{\text{Total fields}}{\text{Total Records}} = 6.31
\]

Number of Incorrect Fields

Correct one field

Correct one field
Confidence Estimation

• 276 records had one or more errors.
• If the least confident field highlighted in a record with one or more errors, an error will be identified 81.9% of the time.
• If field is chosen at random, an error will be identified 29.0% of the time.
• This illustrates the potential for using confidence to direct the users attention to an incorrect field.

Summary

• Synergy of User Interface and Information Extraction Algorithm ensuring confidence integrity of data
• Over 88% reduction of User Actions by Information Extraction alone
• Additional 13% reduction in User Actions due to Correction Propagation
• Confidence Scores effective at identifying incorrect fields.
• IIE in Microsoft Office 2007 ??
End