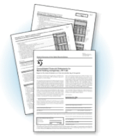
Interactive Information Extraction

Trausti Kristjansson, Aron Culotta, Paul Viola, Andrew
M n

Introduction



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

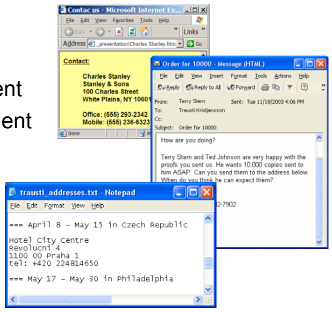
The goal of this work is to reduce the burden on the user to the largest extent possible, while ensuring the integrity of the data.

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



Contac us - Microsoft Internet Ex...
File Edit View Favorites Tools Help
Back
Links
Address _presentation\Charles Stanley.htm
Go
Done

Contact:

Charles Stanley
Stanley & Sons
100 Charles Street
White Plains, NY 10601

Office: (555) 293-2342
Mobile: (555) 236-6323

Order for 10000 - Message (HTML)
File Edit View Insert Format Tools Actions Help
Reply Reply to All Forward

From: Terry Stern Sent: Tue 11/18/2003 4:06 PM
To: Trausti Kristjansson
Cc:
Subject: Order for 10000

How are you doing?

Terry Stern and Ted Johnson are very happy with the proofs you sent us. He wants 10.000 copies sent to him ASAP. Can you send them to the address below. When do you think he can expect them?

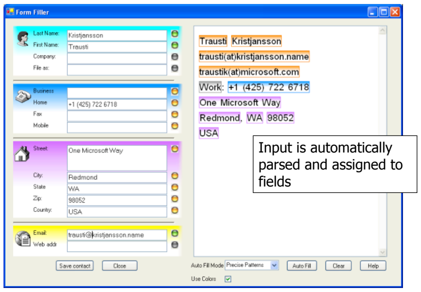
2-7902

```
=== April 8 – May 15 in Czech Republic

Hotel City Centre
Revolucni 4
1100 00 Praha 1
tel: +420 224814650

=== May 17 – May 30 in Philadelphia
```

Demo: Contact Assistant

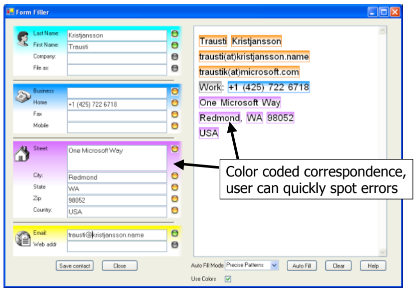


Last Name: Kristjansson
First Name: Trausti
Company:
File as:
Business
Home +1 (425) 722 6718
Fax
Mobile
Street: One Microsoft Way
City: Redmond
State WA
Zip: 98052
Country: USA
Email: trausti@kristjansson.name
Web addr

Trausti Kristjansson
trausti(at)kristjansson.name
traustik(at)microsoft.com
Work: +1 (425) 722 6718
One Microsoft Way
Redmond, WA 98052
USA

Input is automatically parsed and assigned to fields

Data Integrity – Fast Verification



Last Name: Kristjansson
First Name: Trausti
Company:
File as:
Business
Home +1 (425) 722 6718
Fax
Mobile
Street: One Microsoft Way
City: Redmond
State WA
Zip: 98052
Country: USA
Email: trausti@kristjansson.name
Web addr

Trausti Kristjansson
trausti(at)kristjansson.name
traustik(at)microsoft.com
Work: +1 (425) 722 6718
One Microsoft Way
Redmond, WA 98052
USA

Color coded correspondence, user can quickly spot errors

Correction Propagation

- Show live demo

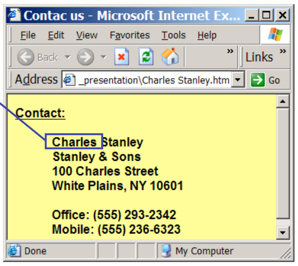
Interactive Information Extraction

- 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


Contac us - Microsoft Internet Ex...
File Edit View Favorites Tools Help
Back
Links
Address
_presentation\Charles Stanley.htm
Go
Done
My Computer

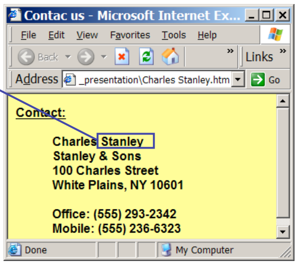
Contact:

Charles Stanley
Stanley & Sons
100 Charles Street
White Plains, NY 10601

Office: (555) 293-2342
Mobile: (555) 236-6323

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


Contac us - Microsoft Internet Ex...
File Edit View Favorites Tools Help
Back
Links
Address
_presentation\Charles Stanley.htm
Go
Done
My Computer

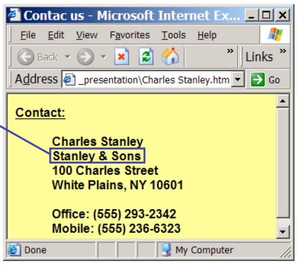
Contact:

Charles Stanley
Stanley & Sons
100 Charles Street
White Plains, NY 10601

Office: (555) 293-2342
Mobile: (555) 236-6323

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


Contac us - Microsoft Internet Ex...
File Edit View Favorites Tools Help
Back
Links
Address
_presentation\Charles Stanley.htm
Go
Done
My Computer

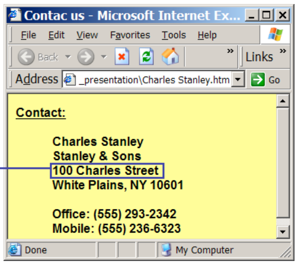
Contact:

Charles Stanley
Stanley & Sons
100 Charles Street
White Plains, NY 10601

Office: (555) 293-2342
Mobile: (555) 236-6323

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

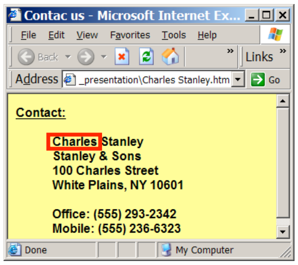

Contac us - Microsoft Internet Ex...
File Edit View Favorites Tools Help
Back
Links
Address _presentation\Charles Stanley.htm
Go
Done
My Computer

Contact:

Charles Stanley
Stanley & Sons
100 Charles Street
White Plains, NY 10601

Office: (555) 293-2342
Mobile: (555) 236-6323

Token Features $f_k(\mathbf{y},t)$


Contac us - Microsoft Internet Ex...
File Edit View Favorites Tools Help
Back
Links
Address _presentation\Charles Stanley.htm
Go
Done
My Computer

Contact:

Charles Stanley
Stanley & Sons
100 Charles Street
White Plains, NY 10601

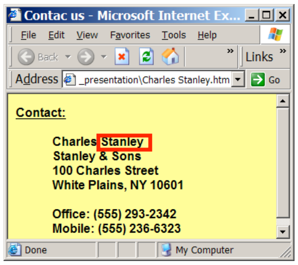
Office: (555) 293-2342
Mobile: (555) 236-6323

- 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 5 Digits
□ Contains Hyphen
□ Enclosed in Brackets

… and 20000 more

Token Features $f_k(\mathbf{y},t)$


Contac us - Microsoft Internet Ex...
File Edit View Favorites Tools Help
Back
Links
Address _presentation\Charles Stanley.htm
Go
Done
My Computer

Contact:

Charles Stanley
Stanley & Sons
100 Charles Street
White Plains, NY 10601

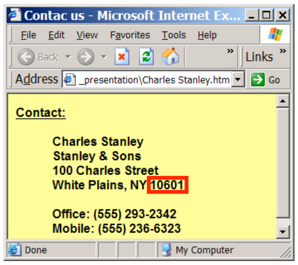
Office: (555) 293-2342
Mobile: (555) 236-6323

- 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 5 Digits
□ Contains Hyphen
□ Enclosed in Brackets

… and 20000 more

Token Features $f_k(\mathbf{y},t)$


Contac us - Microsoft Internet Ex...
File Edit View Favorites Tools Help
Back
Links
Address _presentation\Charles Stanley.htm
Go
Done
My Computer

Contact:

Charles Stanley
Stanley & Sons
100 Charles Street
White Plains, NY 10601

Office: (555) 293-2342
Mobile: (555) 236-6323

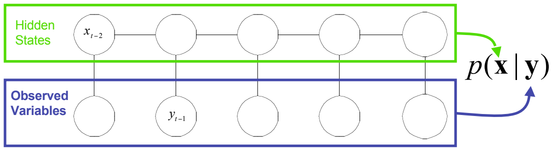
- Features

□ 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 5 Digits
□ Contains Hyphen
□ Enclosed in Brackets

… and 20000 more

Conditional Random Fields

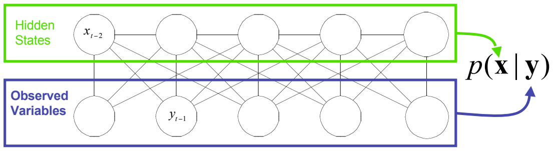
- Conditional Random Fields are *globally normalized* probability models, where hidden variables are *conditioned* on observed variables.



- 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

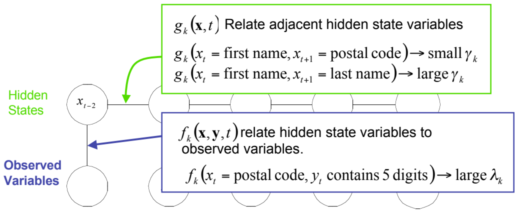
Conditional Random Fields

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

Conditional Random Fields



$g_k(\mathbf{x},t)$ Relate adjacent hidden state variables

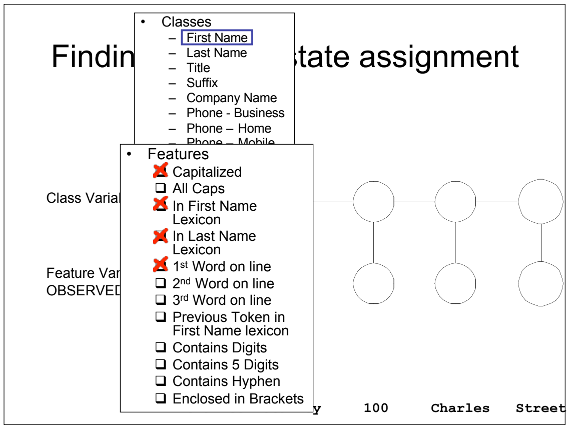
$g_k(x_t = \text{first name}, x_{t+1} = \text{postal code}) \rightarrow \text{small } \gamma_k$
$g_k(x_t = \text{first name}, x_{t+1} = \text{last name}) \rightarrow \text{large } \gamma_k$

$f_k(\mathbf{x},\mathbf{y},t)$ relate hidden state variables to observed variables.

$f_k(x_t = \text{postal code}, y_t \text{ contains 5 digits}) \rightarrow \text{large } \lambda_k$

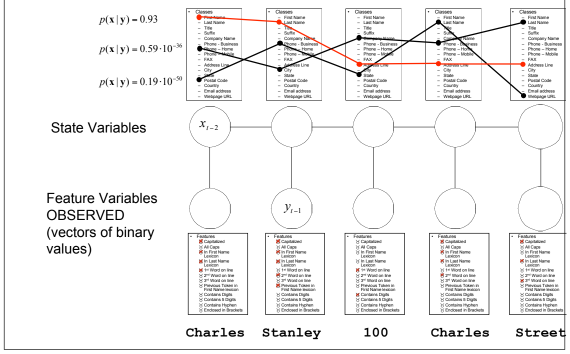
$$p(\mathbf{x}\mid\mathbf{y}) = \frac{1}{Z_0}\exp\left(\sum_{t=1}^{T}\sum_{k}\underline{\gamma_k g_k(\mathbf{x},t)} + \sum_{t=1}^{T}\sum_{k}\underline{\lambda_k f_k(\mathbf{x},\mathbf{y},t)}\right)$$

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



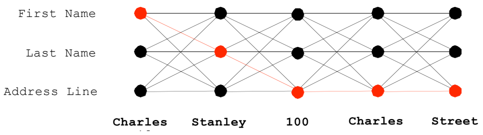
- Classes
  - First Name
  - Last Name
  - Title
  - Suffix
  - Company Name
  - Phone - Business
  - Phone – Home
  - Phone – Mobile
- Features
  - ☒ Capitalized
  - □ All Caps
  - ☒ In First Name Lexicon
  - ☒ In Last Name Lexicon
  - ☒ 1st Word on line
  - □ 2nd Word on line
  - □ 3rd Word on line
  - □ Previous Token in First Name lexicon
  - □ Contains Digits
  - □ Contains 5 Digits
  - □ Contains Hyphen
  - □ Enclosed in Brackets

Finding the best state assignment

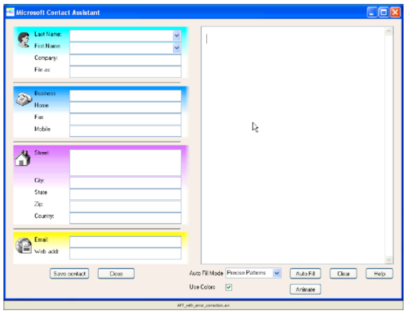


Viterbi used to find best sequence

- Viterbi algorithm may return the sequence of states shown below

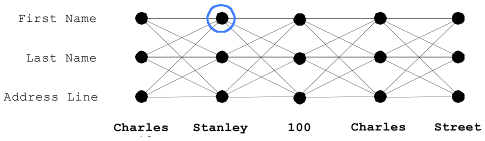


Correction Propagation



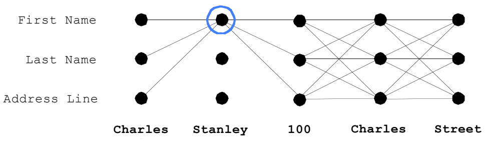
User Correction

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



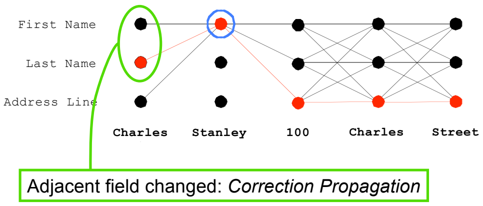
Remove Paths

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



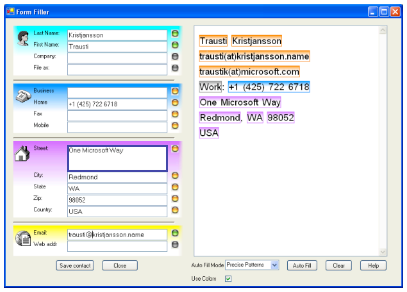
Constrained Viterbi

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



Adjacent field changed: *Correction Propagation*

Indicate Low Confident



Last Name: Kristjansson
First Name: Trausti
Company:
File as:
Business
Home +1 (425) 722 6718
Fax
Mobile
Street: One Microsoft Way
City: Redmond
State WA
Zip: 98052
Country: USA
Email: trausti@kristjansson.name
Web addr

Trausti Kristjansson
trausti(at)kristjansson.name
traustik(at)microsoft.com
Work: +1 (425) 722 6718
One Microsoft Way
Redmond, WA 98052
USA

Auto Fill Mode Precise Patterns
Use Colors ☑

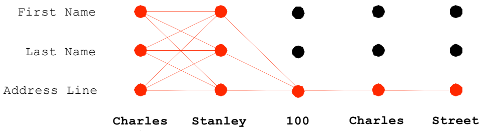
Confidence Estimation

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

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

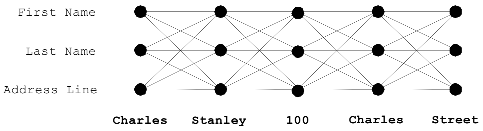
Sum of "agreeing" states sequences

- Paths that "agree" with classification



Sum of all state sequences

- All paths



Evaluation

Standard Metrics

- Standard information retrieval metrics:

	Token Acc.	F1	Precision	Recall
CRF	89.73	87.23	88.24	86.24
MaxEnt	88.43	84.84	85.05	84.95

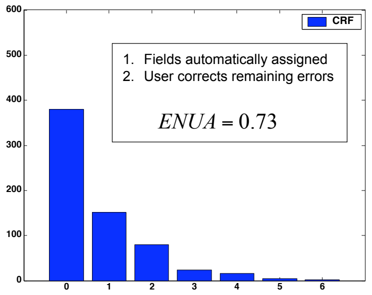
- These metrics don’t relate well to the stated goals, e.g. how much does the system speed up data acquisition.

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.

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

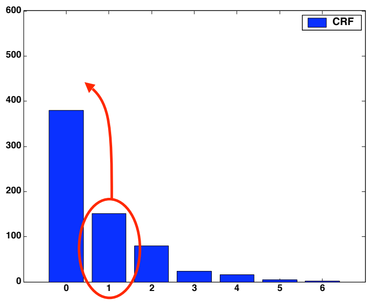
Number of Incorrect Fields



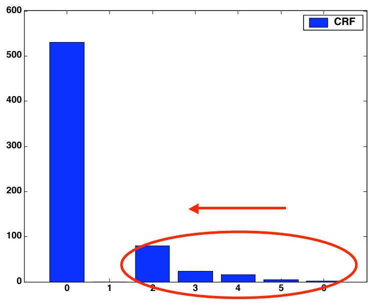
1. Fields automatically assigned
2. User corrects remaining errors

$$ENUA = 0.73$$

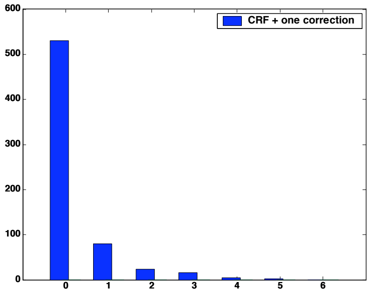
Correct one field



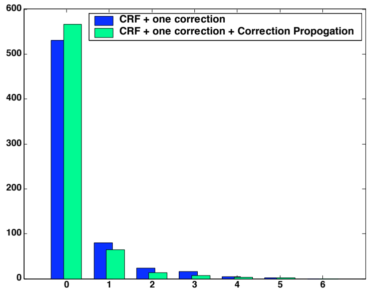
Correct one field



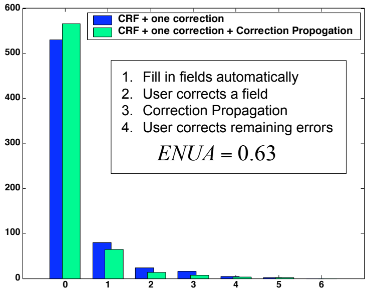
Correct one field



Run correction propagation



Run correction propagation



1. Fill in fields automatically
2. User corrects a field
3. Correction Propagation
4. User corrects remaining errors

$ENUA = 0.63$

Expected Number of User Actions

Model/UI-Model	ENUA	Change
Manual – UIMm	6.31	Baseline
CRF – UIM 1	0.73	-88.4%
CCRF – UIM2	0.63	-93.1%

8.5x

10x

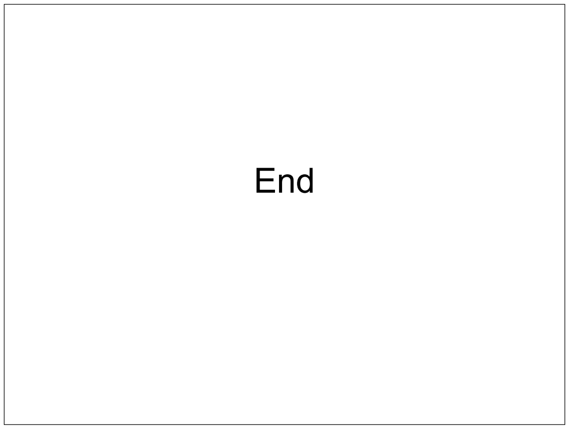
-13.9%

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