Information Extraction

Pedro Szekely
Information Sciences Institute,
USC Viterbi School of Engineering
Agenda

Information extraction classification

Text extraction techniques

Storing extractions in knowledge graphs

myDIG demo

Summary
Astro Teller is the CEO and co-founder of BodyMedia. Astro holds a Ph.D. in Artificial Intelligence from Carnegie Mellon University, where he was inducted as a national Hertz fellow. His M.S. in symbolic and heuristic computation and B.S. in computer science are from Stanford University. His work in science, literature and business has appeared in international media from the New York Times to CNN to NPR.

Dr. Steven Minton - Founder/CTO
Dr. Minton is a fellow of the American Association of Artificial Intelligence and was the founder of the Journal of Artificial Intelligence Research. Prior to founding Fetch, Minton was a faculty member at USC and a project leader at USC’s Information Sciences Institute. A graduate of Yale University and Carnegie Mellon University, Minton has been a Principal Investigator at NASA Ames and taught at Stanford, UC Berkeley and USC.

Frank Huynhbrechts - COO
Mr. Huynhbrechts has over 20 years of

<table>
<thead>
<tr>
<th>Name</th>
<th>Phone</th>
<th>Email</th>
<th>Department</th>
</tr>
</thead>
<tbody>
<tr>
<td>Barto, Andrew G.</td>
<td>(413) 545-2109</td>
<td><a href="mailto:barto@cs.umass.edu">barto@cs.umass.edu</a></td>
<td>CS276</td>
</tr>
<tr>
<td>Brock, Oliver</td>
<td>(413) 577-0334</td>
<td><a href="mailto:oli@cs.umass.edu">oli@cs.umass.edu</a></td>
<td>CS246</td>
</tr>
<tr>
<td>Clarke, Lori A.</td>
<td>(413) 545-1328</td>
<td><a href="mailto:clarke@cs.umass.edu">clarke@cs.umass.edu</a></td>
<td>CS304</td>
</tr>
<tr>
<td>Cohen, Paul R.</td>
<td>(413) 545-3638</td>
<td><a href="mailto:cohen@cs.umass.edu">cohen@cs.umass.edu</a></td>
<td>CS278</td>
</tr>
</tbody>
</table>

Professor.

Computational neuroscience, reinforcement learning, adaptive motor control, artificial neural networks, adaptive and learning control, motor development.

Berger, Emery D. | (413) 577-4211 | emcyr@cs.umass.edu | CS344        |
Assistant Professor.

Professor.

Software verification, testing, and analysis; software architecture and design.

Cohen, Paul R.  | (413) 545-3638 | cohen@cs.umass.edu | CS278        |
Professor.

Planning, simulation, natural language, agent-based systems, intelligent data analysis, intelligent user interfaces.
Scope

Web site specific

Genre specific (e.g., forums)

Wide, non-specific
Pattern Complexity

E.g., word patterns

Closed set

U.S. states

He was born in Alabama...
The big Wyoming sky...

Regular set

U.S. phone numbers

Phone: (413) 545-1323

The CALD main office can be reached at 412-268-1299

Complex pattern

U.S. postal addresses

University of Arkansas
P.O. Box 140
Hope, AR 71802

Headquarters:
1128 Main Street, 4th Floor
Cincinnati, Ohio 45210

Ambiguous patterns, needing context and many sources of evidence

Person names

...was among the six houses sold by Hope Feldman that year.
Pawel Opalinski, Software Engineer at WhizBang Labs.

“You don’t wanna miss out on ME :) Perfect lil booty Green eyes Long curly black hair Im a Irish, Armenian and Filipino mixed princess :) ❤ Kim ❤ 7o7~7two7~7four77 ❤ HH 80 roses ❤ Hour 120 roses ❤ 15 mins 60 roses”
Practical Considerations

How good (precision/recall) is necessary?
High precision when showing extractions to users
High recall when used for ranking results

How long does it take to construct?
Minutes, hours, days, months

What expertise do I need?
None (domain expertise), patience (annotation), simple scripting, machine learning guru

What tools can I use?
Many …
Information Extraction Process

Segmentation

Data Extraction
Information Extraction Process

Segmentation

Data Extraction
Information Extraction Process

Segmentation

Data Extraction

Name: Legacy Ventures Intl, Inc.

Stock: LGYV

Date: 2017-07-14

Market Cap: 391,030
Segmentation

Semi-structured extraction

Table extraction

Main content identification

Custom regular expressions
Segmentation

Semi-structured extraction
Table extraction
Main content identification
Custom regular expressions

Text segments
Text Extraction Techniques

Glossary

Regular expressions

Natural language rules

Named entity recognition

Sequence labeling (Conditional Random Fields)
Glossary Extraction
Glossary Extraction

Simple
list of words or phrases to extract

Challenges
Ambiguity: Charlotte is a name of a person and a city
Colloquial expressions: “Asia Broadband, Inc.” vs “Asia Broadband”

Research
Improving precision of glossary extractions using context
Creating/extending glossaries automatically
Regex Extraction
Extraction Using Regular Expressions

Too difficult for non-programmers

regex for North American phone numbers:

```
^(?:(?:(?:\+?1\s*\(?:[^-]\s*)\)?\s*(?:\d*([2-9]\s*[02-9]|\s*[3-9]\s*[02-9]|\s*[2-9]\s*[02-9]\s*[02-9]|\s*[0-9]{4})(?:(?:\#|x\.|\s*ext\.|\s*extension)\s*(?:\d+))?)$)
```

Brittle and difficult to adapt to unusual domains

unusual nomenclature and short-hands

obfuscation
NLP Rule-Based Extraction
NLP Rule-Based Extraction

Tokenization

Pattern Matching
Tokenization

My name is Pedro

310-822-1511

♥Candy♥ is here

My name is Pedro

310-822-1511

310 - 822 - 1511

♥Candy♥ is here

♥Candy♥ is here
Token Properties

Surface properties
Literal, type, shape, capitalization, length, prefix, suffix, minimum, maximum

Language properties
Part of speech tag, lemma, dependency
Token Types

Create Word Token

- Optional
- Part of output
- Match lemma
- Alphanumeric

Words:

Part of speech:
- Noun
- Pronoun
- Proper noun
- Determiner
- Symbol
- Adjective
- Conjunction
- Verb
- Pre/post-position
- Adverb
- Particle
- Interjection

Capitalization:
- Exact
- Lower
- Upper
- Title
- Mixed

Length 1:  | Length 2:  | Length 3:  
Prefix:  | Suffix:  | Not in vocabulary

Create Shape Token

- Optional
- Part of output

Shape:
Enter shapes such as ddd, XXXX, XX, d is for digits and x for letter, X for capital letter.

Part of speech:
- Noun
- Pronoun
- Proper noun
- Determiner
- Symbol
- Adjective
- Conjunction
- Verb
- Pre/post-position
- Adverb
- Particle
- Interjection

Prefix:  | Suffix:  | Cancel  Save

Create Number Token

- Optional
- Part of output

Numbers:

- Length 1:  | Length 2:  | Length 3:  
Min:  | Max:  | Cancel  Save

Create Punctuation Token

- Optional
- Part of output

Punctuation Symbols:
- ,  | !  | <  
- .  | (  | )  >  
- ;  | [  | ]  =  
- ?  | {  | }  
- ~  | \  | 
- :  | `  | *  
- '  |  \  | $  
- +  | -  | @  
- _  | ^  | @  

Cancel  Save
Patterns

Pattern := Token-Spec

[Token-Spec] Optional

Token-Spec + One or more

Token-Spec Pattern
Positive/Negative Patterns

Positive
Generate candidates

Negative
Remove candidates
Output overlaps positive candidates
Positive/Negative Patterns

General
- **Positive**
  - Generate candidates

Specific
- **Negative**
  - Remove candidates
  - Output overlaps positive candidates
Rule-based matching

spaCy features a rule-matching engine that operates over tokens, similar to regular expressions. The rules can refer to token annotations and flags, and matches support callbacks to accept, modify and/or act on the match. The rule matcher also allows you to associate patterns with entity IDs, to allow some basic entity linking or disambiguation.

Here’s a minimal example. We first add a pattern that specifies three tokens:

1. A token whose lower-case form matches "hello"
2. A token whose is_punct flag is set to True
3. A token whose lower-case form matches "world"

Once we’ve added the pattern, we can use the matcher as a callable, to receive a list of (ent_id, start, end) tuples.

```python
from spacy.matcher import Matcher
from spacy.attrs import IS_PUNCT, LOWER

matcher = Matcher(nlp.vocab)
matcher.add_pattern("HelloWorld", [[{LOWER: "hello"}, {IS_PUNCT: True}], {LOWER: "world"}])
```
Advantages/Disadvantages

Advantages
Easy to define
High precision
Recall increases with number of rules

Disadvantages
Text must follow strict patterns
NLP Rule-Based Extraction

Tokenization for unusual domains
-tokenize on white-space, punctuation and emojis

Token properties
-literal, part of speech tag, lemma, in/out of dictionary
-dependency parsing relationships (advanced)
type (alphanumerical, alphabetic, numeric)
-shape (pattern of digits and characters), capitalization, prefix and suffix
-number of characters, range (numbers)

Pattern
-Sequence of required/optional tokens
-positive and negative patterns
Named-Entity Recognizers
Named Entity Recognizers

Machine learning models
people, places, organizations and a few others

SpaCy
complete NLP toolkit, Python (Cython), MIT license
code: https://github.com/explosion/spaCy
demo: http://textanalysisonline.com/spacy-named-entity-recognition-ner

Stanford NER
part of Stanford’s NLP software library, Java, GNU license
code: https://nlp.stanford.edu/software/CRF-NER.shtml
demo: http://nlp.stanford.edu:8080/ner/process
Entity recognition

spaCy features an extremely fast statistical entity recognition system, that assigns labels to contiguous spans of tokens. The default model identifies a variety of named and numeric entities, including companies, locations, organizations and products. You can add arbitrary classes to the entity recognition system, and update the model with new examples.

The standard way to access entity annotations is the doc.ents property, which produces a sequence of Span objects. The entity type is accessible either as an integer ID or as a string, using the attributes ent.label and ent.label_. The Span object acts as a sequence of tokens, so you can iterate over the entity or index into it. You can also get the text form of the whole entity, as though it were a single token. See the API reference for more details.

You can access token entity annotations using the token.ent_iob and token.ent_type attributes. The token.ent_iob attribute indicates whether an entity starts, continues or ends on the tag (In, Begin, Out).
2 April 2016: Nigeria: NLC Pledges Support for EFCC Anti-Corruption War By Ronald Mutum

The Nigeria Labour Congress (NLC) has thrown its weight in support of the Economic and Financial Crimes Commission (EFCC) anti-corruption campaign. The president of the workers’ union, Ayuba Wabba, gave the Union’s unalloyed support in the fight against corruption during a visit to the chairman of the EFCC, Ibrahim Magu his Abuja office. A statement yesterday from the EFCC spokesman Wilson Uwuja as
Advantages/Disadvantages

Advantages
Easy to use
Tolerant of some noise
Easy to train

Disadvantages
Performance degrades rapidly for new genres, language models
Requires hundreds to thousands of training examples
Conditional Random Fields
### Discriminative Vs. Generative

- **Generative Model**: A model that generates observed data randomly
- **Naïve Bayes**: once the class label is known, all the features are independent
  \[ p(y, x) = p(y) \prod_{k=1}^{K} p(x_k | y) \]
- **Discriminative**: Directly estimate the posterior probability; Aim at modeling the “discrimination” between different outputs
- **MaxEnt classifier**: linear combination of feature function in the exponent,
  \[ p(y|x) = \frac{1}{Z(x)} \exp \left\{ \sum_{k=1}^{K} \theta_k f_k(y, x) \right\} \]

Both generative models and discriminative models describe distributions over \((y, x)\), but they work in different directions.
Discriminative Vs. Generative

$p(y, x)$

Naive Bayes

SEQUENCE

HMMs

GENERAL GRAPHS

Generative directed model

$p(y | x)$

Logistic Regression

SEQUENCE

Linear-chain CRFs

GENERAL GRAPHS

General CRFs

- = observable

= unobservable

Slide by Daniel Khashabi
Chain CRFs

- Each potential function will operate on pairs of adjacent label variables

\[
p(y|x, \lambda) = \frac{1}{Z(x)} \exp \left( \sum_j \lambda_j F_j(y, x) \right)
\]

\[
F_j(y, x) = \sum_{i=1}^{f_j} f_j(y_{i-1}, y_i, x, i),
\]

- Parameters to be estimated,

\[\lambda_j\]
Chain CRF

- We can change it so that each state depends on more observations

- Or inputs at previous steps

- Or all inputs
# Modeling Problems With CRF

<table>
<thead>
<tr>
<th>i</th>
<th>X1 (word)</th>
<th>X2 (capitalized)</th>
<th>X3 (POS Tag)</th>
<th>Y (entity)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>My</td>
<td>1</td>
<td>Possessive Pron</td>
<td>Other</td>
</tr>
<tr>
<td>2</td>
<td>name</td>
<td>0</td>
<td>Noun</td>
<td>Other</td>
</tr>
<tr>
<td>3</td>
<td>is</td>
<td>0</td>
<td>Verb</td>
<td>Other</td>
</tr>
<tr>
<td>4</td>
<td>Pedro</td>
<td>1</td>
<td>Proper Noun</td>
<td>Person-Name</td>
</tr>
<tr>
<td>5</td>
<td>Szekely</td>
<td>1</td>
<td>Proper Noun</td>
<td>Person-Name</td>
</tr>
</tbody>
</table>
## Modeling Problems With CRF

<table>
<thead>
<tr>
<th>i</th>
<th>X1 (word)</th>
<th>X2 (capitalized)</th>
<th>X3 (POS Tag)</th>
<th>Y (entity)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>My</td>
<td>1</td>
<td>Possessive Pron</td>
<td>Other</td>
</tr>
<tr>
<td>2</td>
<td>name</td>
<td>0</td>
<td>Noun</td>
<td>Other</td>
</tr>
<tr>
<td>3</td>
<td>is</td>
<td>0</td>
<td>Verb</td>
<td>Other</td>
</tr>
<tr>
<td>4</td>
<td>Pedro</td>
<td>1</td>
<td>Proper Noun</td>
<td>Person-Name</td>
</tr>
<tr>
<td>5</td>
<td>Szekely</td>
<td>1</td>
<td>Proper Noun</td>
<td>Person-Name</td>
</tr>
</tbody>
</table>

Other common features:
- lemma, prefix, suffix, length
# Modeling Problems With CRF

<table>
<thead>
<tr>
<th>i</th>
<th>X1 (word)</th>
<th>X2 (capitalized)</th>
<th>X3 (POS Tag)</th>
<th>Y (entity)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>My</td>
<td>1</td>
<td>Possessive Pron</td>
<td>Other</td>
</tr>
<tr>
<td>2</td>
<td>name</td>
<td>0</td>
<td>Noun</td>
<td>Other</td>
</tr>
<tr>
<td>3</td>
<td>is</td>
<td>0</td>
<td>Verb</td>
<td>Other</td>
</tr>
<tr>
<td>4</td>
<td>Pedro</td>
<td>1</td>
<td>Proper Noun</td>
<td>Person-Name</td>
</tr>
<tr>
<td>5</td>
<td>Szekely</td>
<td>1</td>
<td>Proper Noun</td>
<td>Person-Name</td>
</tr>
</tbody>
</table>

feature functions  $f_j(x, y_{i-1}, y_i, i)$
Advantages/Disadvantages

**Advantages**
- Expressive
- Tolerant of noise
- Stood test of time
- Software packages available

**Disadvantages**
- Requires feature engineering
- Requires thousands of training examples
Open Information Extraction
http://openie.allenai.org/

Example Queries: *
What kills bacteria?
Who built the Pyramids?
What did Thomas Edison invent?
What contains antioxidants?

Typed Example Queries: *
What countries are located in Africa?
What actors starred in which films?
What is the symbol of which country?
What foods are grown in which countries?
What drug ingredients has the FDA approved?

Argument 1: Relation: Argument 2: Corpus: [ ] All [ ]

AI2 proudly announces the launch of Semantic Scholar, an AI-based academic search engine.

To learn more about Open IE, watch our YouTube video!

Powered by ReVerb, our Open Information Extractor, yielding over 5 billion extractions from over a billion web pages.

**New Open IE 4.0**, the successor to ReVerb and Ollie, has been released. Download it from GitHub!

Publications:
- Search Needs a Shake-up (Nature 2011)
- Open Information Extraction (IJCAI 2011)
- Ollie (EMNLP 2012)
- Reverb (EMNLP 2011)
- TextRunner (IJCAI 2007)

Public resources based on Open IE:
- 18 million question-paraphrases (Fader et al. ACL 2013)
# Practical IE Technologies

<table>
<thead>
<tr>
<th>Effort</th>
<th>Glossary</th>
<th>Regex</th>
<th>NLP Rules</th>
<th>Semi-Structured</th>
<th>CRF</th>
<th>NER</th>
<th>Table</th>
</tr>
</thead>
<tbody>
<tr>
<td>assemble glossary</td>
<td>hours</td>
<td>hours</td>
<td>minutes</td>
<td>O(1000) annotations</td>
<td>zero</td>
<td>O(10) annotations</td>
<td></td>
</tr>
<tr>
<td>Expertise</td>
<td>minimal</td>
<td>high, programmer</td>
<td>low</td>
<td>minimal</td>
<td>low-medium</td>
<td>zero</td>
<td>minimal</td>
</tr>
<tr>
<td>Precision</td>
<td>medium (ambiguity)</td>
<td>high</td>
<td>high</td>
<td>high</td>
<td>medium-high</td>
<td>medium-high</td>
<td>high</td>
</tr>
<tr>
<td>Recall</td>
<td>medium (formatting)</td>
<td>low f(# regex)</td>
<td>medium f(# rules)</td>
<td>high</td>
<td>medium</td>
<td>medium</td>
<td>high</td>
</tr>
</tbody>
</table>
how to represent KGs?
KG Definition

a directed, labeled multi-relational graph representing facts/assertions as triples

(h, r, t) head entity, relation, tail entity
(s, p, o) subject, predicate, object
Simplest Knowledge Graph

Entities

mentions

mentions

mentions

LGY V

Legacy Ventures International Inc

Damn Good Penny Stocks

Easiest to build
Simple, But Useful KG

Entities + properties

stock-ticker

company

promoter

LGY

Legacy Ventures International Inc

Damn Good Penny Stocks

“Easy” to build
Semantic Web KG (RDF/OWL)

Entities + properties + classes

Company

Legacy Ventures International Inc

Damn Good Penny Stocks

stock-ticker

is-a

promoter

Very hard to build

Kejriwal, Szekely
“Ideal” KG

Entities + properties + classes + qualifiers

Compan

y

is-a

is-a

Damn Good Penny Stocks

is-a

Legacy Ventures International Inc

stock-ticker

LGY

promoter

source

start-date

June 2017

stockreads.com

Very very very hard to build
"More Ideal" KG

Entities + properties + provenance + confidence + qualifiers

"Not so hard" to build
Where to Store KGs?
Serializing Knowledge Graphs

Resource Description Framework (RDF)
Database (triple store): AllegroGraph, Virtuoso,
Query: SPARQL (SQL-like)

Key-Value, Document Stores
Data model: Node-centric
Databases: Hbase, MongoDB, Elastic Search, …
Query: filters, keywords, aggregation (no joins)

Graph Databases
Data model: graph
Databases: Neo4J, Cayley, MarkLogic, GraphDB, Titan, OrientDB, Oracle, …
Query: GraphQL, Gremlin, Cypher
Popularity Ranking Of Graph Databases

DB-Engines Ranking of Graph DBMS

© August 2017, DB-Engines.com

https://db-engines.com/en/ranking_trend/graph+dbms
ElasticSearch, MongoDB & Neo4J Have Wide Adoption

DB-Engines Ranking

https://db-engines.com/en/ranking_trend/graph+dbms
myDIG: A KG Construction Toolkit
Python, MIT license, https://github.com/usc-isri-i2/dig-etl-engine

Enable end-users to construct domain-specific KGs
end users from 5 government orgs constructed KGs in less than one day

Suite of extraction techniques
semi-structured HTML pages, glossaries, NLP rules, NER, tables (coming soon)

KG includes provenance and confidences
enable research to improve extractions and KG quality

Scalable
runs on laptop (~100K docs), cluster (> 100M docs)

Robust
Deployed to many law enforcement agencies

Easy to install
Docker deployment with single “docker compose up” installation
myDIG Demo
Summary

Partition pages into segments
Select technology based on segment features
Do knowledge graph completion (next topic)
Choose representation based on application demands