



GOVERNMENT
USERS CONFERENCE

JUNE 8-9
2010
CHANTILLY, VA

A Gentle Introduction to Entity Extraction

Brandon Mensing
Software Engineer
Basis Technology Corp.



Human Language Technology From Arabia to Afghanistan

Agenda

- Some definitions
- Why extract entities?
- Approaches to entity extraction
- Evaluating entity extraction results

What is an Entity?

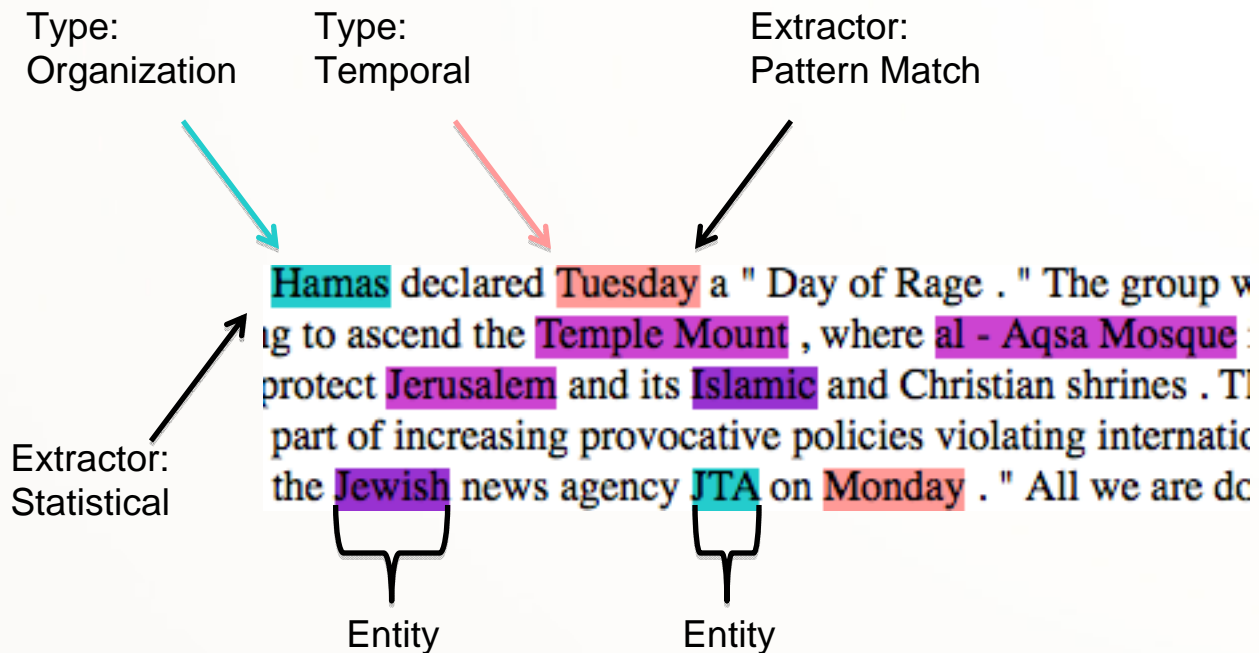
- A substring of text that fits into a predefined category of interest.
 - Especially People, Locations and Organizations

Hamas declared **Tuesday** a " Day of Rage ." The group w
ig to ascend the **Temple Mount** , where **al - Aqsa Mosque** :
protect **Jerusalem** and its **Islamic** and Christian shrines . Tl
part of increasing provocative policies violating internatic
the **Jewish** news agency **JTA** on **Monday** . " All we are dc

Definitions

- Entity
 - The text of interest. Also called a 'mention'.
 - Examples: "Obama", "Nevada", "Google"
- Entity Type
 - A category of entities.
 - Examples: Person, Location, Organization
- Entity Extractor
 - A system that identifies entities and their entity types.

Visual Definitions



Why do you want Entities?

- Faceted search
- Aid in Machine Translation
- Intelligence gathering
- Multi-document analysis
- Fact extraction
- Relationship extraction

REX

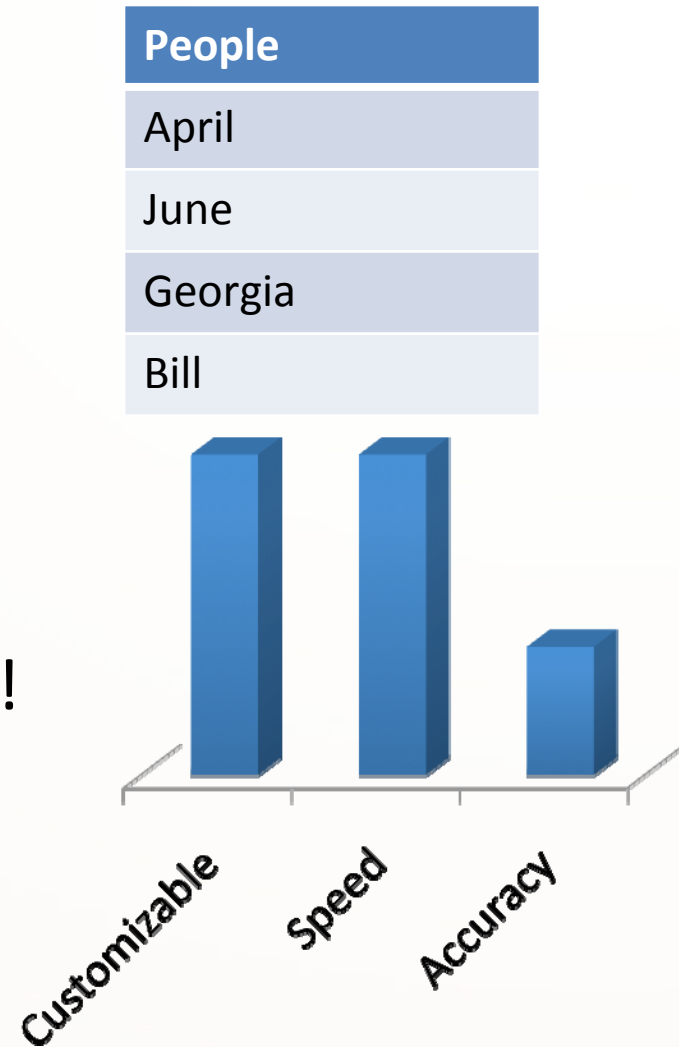
- Part of Rosette 7
- Multifaceted approach
- Highly customizable
- Tuned for speed and accuracy

Approaches to Entity Extraction

- Really simple
- Kinda simple
- Not so simple

Really Simple

- Just use a list!
- Pros:
 - Easy to customize.
 - Fast, really fast.
- Cons:
 - No flexibility
 - Can easily be very wrong!



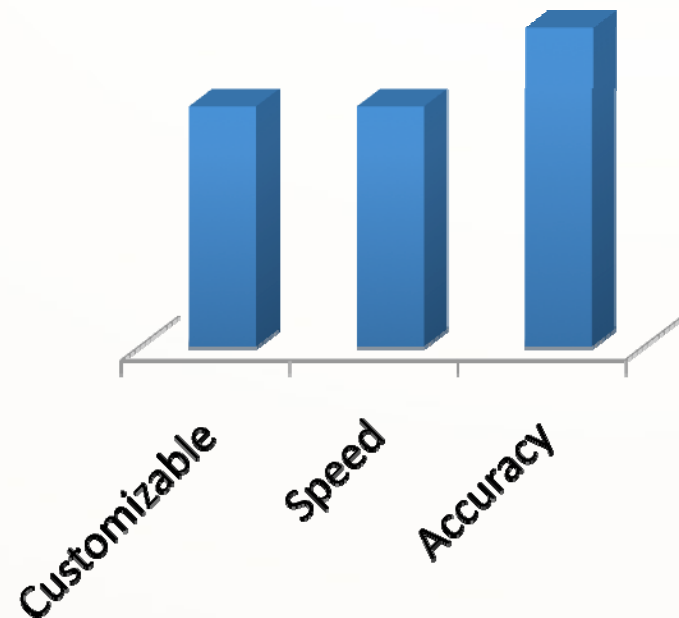
REX's Exact Match Extractor

- Fast tree-based search algorithm
- Simple interface for customization

Kinda Simple

- Define patterns to find entities
- Pros:
 - Somewhat adaptable
 - User configurable
- Cons:
 - Difficult to use
 - Slow
 - Still rules-based

Date	##/##/####
Time	##:## [AM PM]

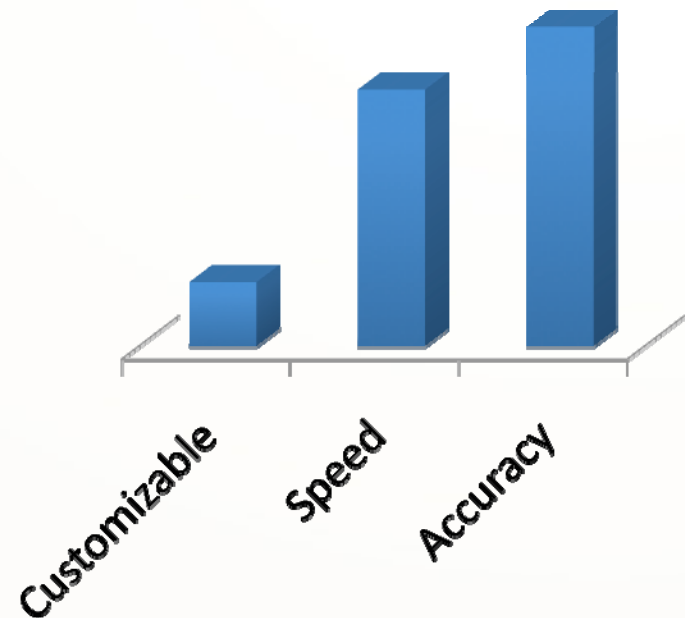


REX's Pattern Match Extractor

- Based on standard Regular Expressions
- Uses an efficient implementation of regular expression search
- Easy to customize and adapt

Not So Simple

- ‘Teach’ the computer how to recognize entities without using rules
- Pros:
 - The most adaptable.
 - Captures entities without being told exactly how to find them.
- Cons:
 - Not (usually) customizable



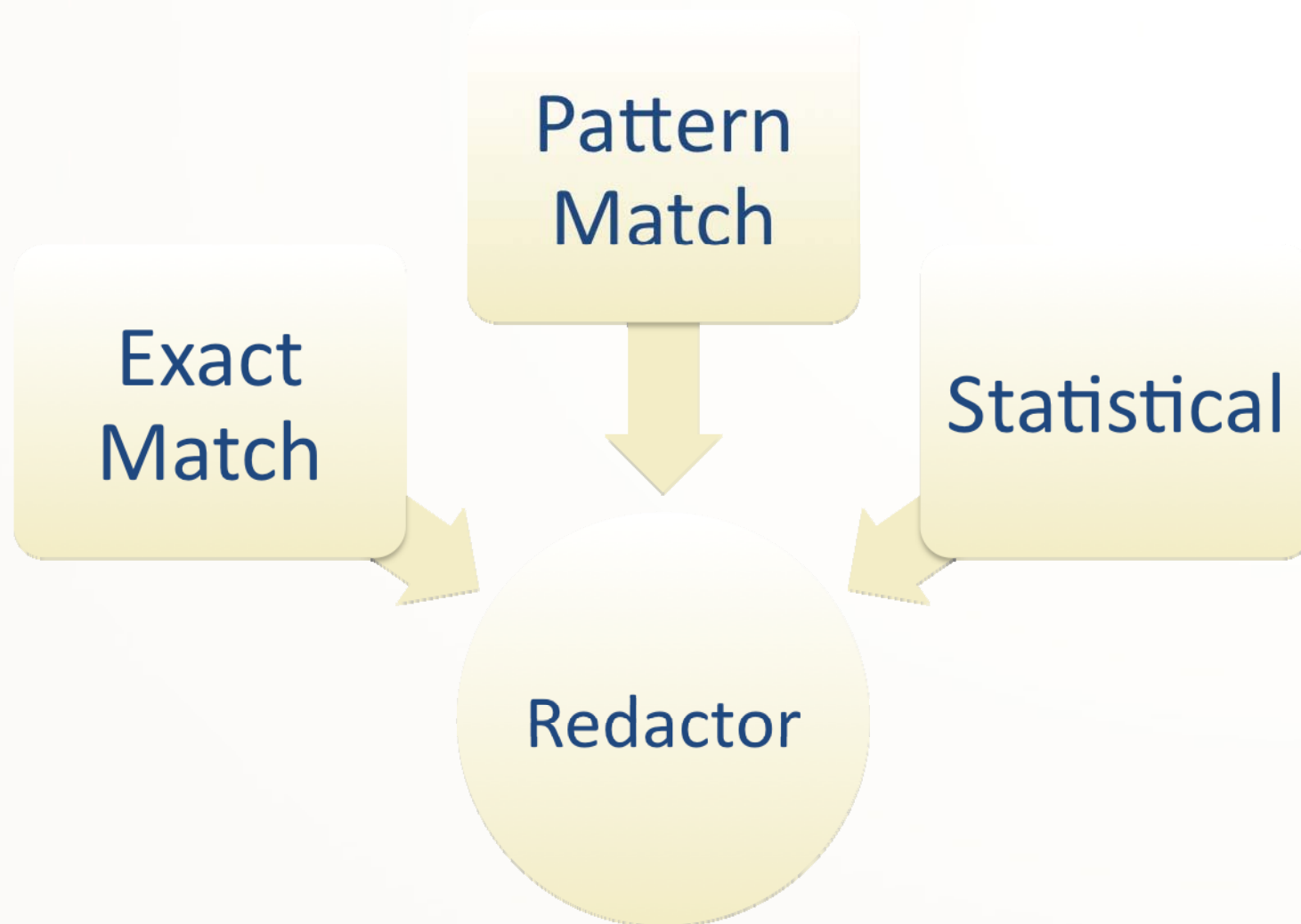
REX's Statistical Extractor

- State of the art learning algorithm
- Well-tuned for various data domains
- Fast
- Training time is so small that field training is possible
 - In other words, it's customizable!
 - Not part of standard package, but please talk to us if you are interested.

Customizable Redaction

- 3 extractors could conflict
- REX's redactor resolves
 - Customizable settings
 - Attend the tutorial for more info

REX



Evaluation

- What are we comparing?
 - Human-annotated evaluation data
 - Machine-annotated evaluation data
- How do you know how good it is?
 - Precision, Recall and F (Accuracy)

Evaluation Definitions

- Precision
 - For any entity output by the machine, it is *precise* only if the human annotated it as well.
- Recall
 - For any entity annotated by a human, that entity is *recalled* only if the machine output it as well.

Evaluation Definitions (Cont.)

- F
- The *harmonic mean* of Precision and Recall.
 - (We use $\beta = 1.0$)

$$F = (1 + \beta^2) \bullet \frac{\textit{precision} \bullet \textit{recall}}{\textit{precision} + \textit{recall}}$$

Evaluation: Precision

- The final precision score is the ratio of correct entities to the number machine-extracted entities.

Precision

Dorothy used the burlap sac to carry her groceries out of **Wal-Mart**.

Dorothy used the **burlap sac** to carry her groceries out of **Wal-Mart**.

Evaluation: Recall

- The final recall score is the ratio of correct entities to the number of human-annotated entities.

Recall

Dorothy used the burlap sac to carry her groceries out of Wal-Mart.

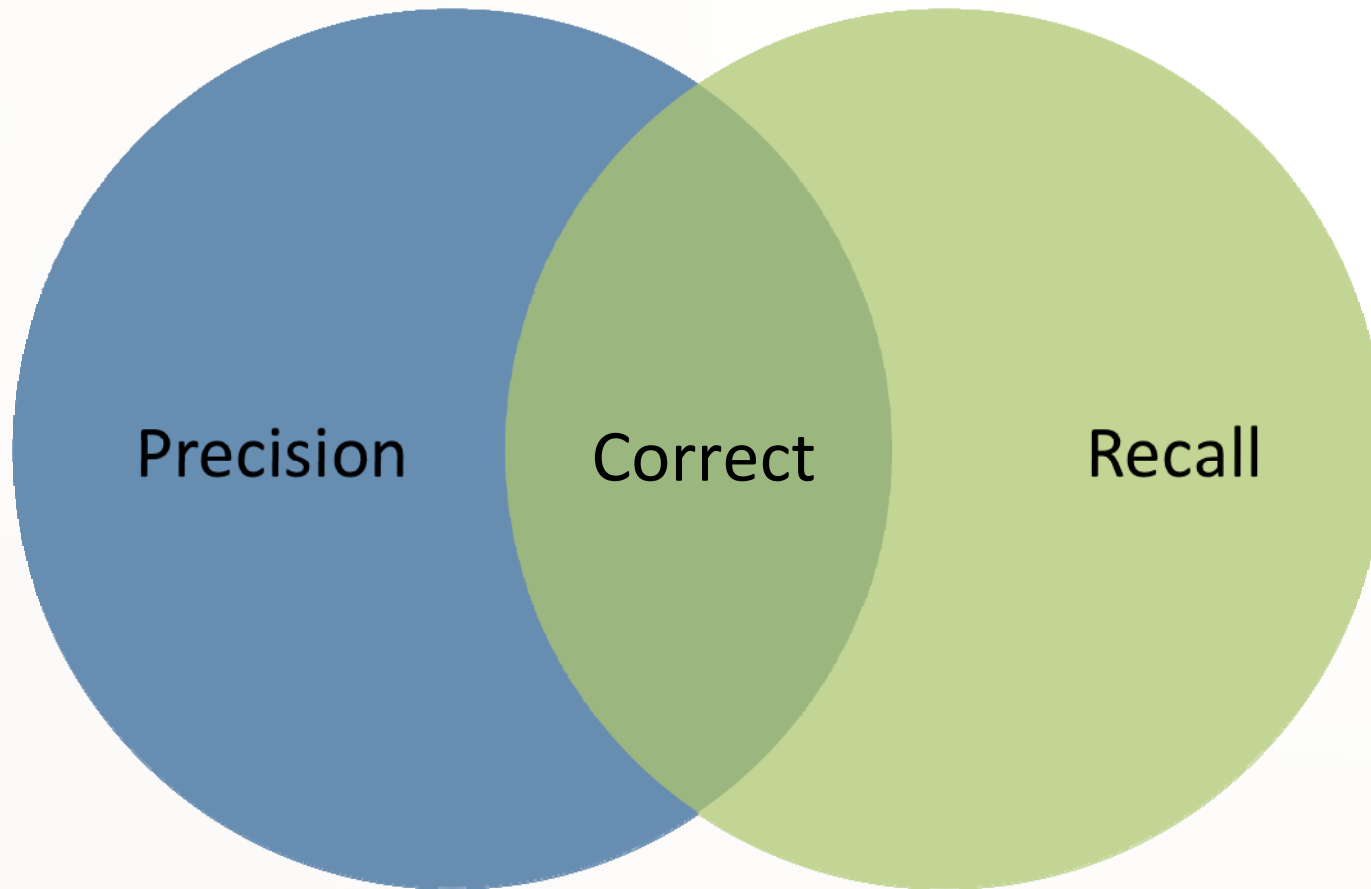
Dorothy used the burlap sac to carry her groceries out of Wal-Mart.

F Keeps it Fair

- Precision and Recall are both important to meet variable use cases
- F provides enforcement of that fact

$$F = (1 + \beta^2) \bullet \frac{\textit{precision} \bullet \textit{recall}}{\textit{precision} + \textit{recall}}$$

F Defined Visually



F Example

- $P = 0.1$
- $R = 0.99$
- ~~Average = 0.595~~
- $F = 0.182$

Real F

- $P = 0.88$
- $R = 0.88$
- ~~Average = 0.880~~
- $F = 0.880$
 - (REX Statistical Extraction for English)

Questions?



For More Information

- Visit www.basistech.com
- Write to info2010@basistech.com
- Call 617-386-2090 or 800-697-2062

Thank you!

