

A Gentle Introduction to Entity Extraction and the Rosette Entity Extractor

次世代サーチ・テクノロジー・フォーラム東京
2010年4月22日

Benjamin Douglas, Principal Engineer

www.basistech.com



Entities and Entity Extraction

Entities and Entity Extraction

- Entities are people, places, organizations, and terms of interest
- Entity extraction is the process of automatically finding and categorizing entities in unstructured text

Uses of Entity Extraction



BASIS
TECHNOLOGY

Uses of Entity Extraction

- Search result navigation
- Search relevance
- Rough summarization
- Other examples

Search Result Navigation

The image shows a screenshot of a Japanese e-commerce search results page. A search bar at the top left contains the keyword 'モーニング娘' (Morning Musume). Below it, there are filters for 'ジャンル' (Genre) and '配送' (Shipping). A large red-bordered box is overlaid on the page, titled 'ジャンルを指定して絞り込む' (Filter by specifying genre). This box contains a list of genres with their respective item counts: CD・DVD・楽器 (7,812), インテリア・寝具・収納 (5), おもちゃ・ホビー・ゲーム (217), and キッチン・日用品雑貨・文具 (39). The background page shows various product listings, including 'シンクビー!', '【即日発送】お洒落さんに', 'わけあり超得 青見オレンジ', and '▲100円!▲ 毎週替わる! 感想15,400件'. At the bottom, there are additional filters like '表示方法' (Display method) and '絞り込み条件' (Filtering conditions).

検索条件

■検索キーワード
モーニング娘

をすべて含む

商品名、商品番号で探す
-検索条件を追加する

キーワード
を除外

価格 価格帯で絞り込む
円～ 円
商品検索

要日配送で絞り込む

要日配送で絞り込む

ジャンルを指定して絞り込む

CD・DVD・楽器 (7,812)
インテリア・寝具・収納 (5)
おもちゃ・ホビー・ゲーム (217)
キッチン・日用品雑貨・文具 (39)

ジャンルを指定して絞り込む

ジャンルを指定して絞り込む

表示方法: 写真付き一覧 | 写真なし一覧 | ウィンドウショッピング

絞り込み条件: 在庫あり 送料込み カードOK ギフト対応可 感想あり 絞り

要日配送可能エリア 絞り込み条件を追加する

他のキーワード: [モーニング娘 loveマシーン](#) 買い物可能 共同購入 スーパーオー

Search Relevance

- Co-reference, boosting, filtering

The image shows a screenshot of a Twitter search interface. The search bar contains the text "Demi Moore". Below the search bar, there are two main sections: "User" and "Person/Concept". The "Person/Concept" section is highlighted with a green background and contains a dropdown menu with the following items: "Demi Moore (26)", "Ashton Kutcher (10)", "Demi (3)", and "...More". A callout box with a black border and a green header "Person/Concept" is positioned over the dropdown menu, showing the same items. The main search results area shows a list of tweets, including one from [iQonz](#) and another from [RickAlanRoss](#) mentioning "Kabbalah pals Madonna and Demi Moore".

Search: Demi Moore

Welcome

Browse

User

mrskutcher (6)

DemilyShelton (2)

MandyLovely (2)

...More

Person/Concept

Demi Moore (26)

Ashton Kutcher (10)

Demi (3)

...More

1-10 of 59

[iQonz](#): OOooooo, wait, I v
man, I'm expecting anoth
2009-05-12T22:35:40Z - Re

[MY NBA PLAYOFFS](#): ... Deborah Gibson, Deborah Perez, Debra Messing, Debra
Wilson, Demi Lovato, Demi Moore, Denis Leary, Denise Ric.. <http://bit.ly/2bNsG>
2009-06-12T21:44:00Z - [Reply](#) - [View Tweet](#)

[RickAlanRoss](#): **Kabbalah** pals **Madonna** and **Demi Moore** compare notes on "boy toys." Is
dating younger men becoming a religious rite?. See www.cultnews.net
2009-05-02T12:00:00Z - [Reply](#) - [View Tweet](#)

2 3 4 5 ▶

re, man-o-

SORT ▾

Rough Summarization

郵貯上限2千万円再調整、首相「了解していない」

鳩山首相は25日夕、亀井郵政改革相と原口総務相が発表した郵政改革法案の「最終案」に関し、「まだ議論すべきところが残っている。強力な案であることは間違いなが、閣議の場などで調整していくことが必要だ」と述べ、同案の骨格であるゆうちょ銀行への預入限度額引き上げの是非などについて、閣内で議論をやり直す考えを表明した。

26日の閣議で亀井氏らに指示する意向だが、担当閣僚が発表した政府の重要政策について、首相が仕切り直しを指示するのは異例の事態だ。

首相は、亀井氏から23日に発表内容の報告を電話で受けたことを明らかにしたうえで、「閣内で議論する前にあたかもすべて決まったかのように発言された。調整前に発表したことはまずかった」と述べ、亀井氏らの対応に問題があったと指摘した。亀井氏が首相の了解を得て発表したと説明していることについても、

PERSON:

亀井 (7)

原口 (2)

亀井郵政改革相 (1)

原口総務相 (1)

鳩山首相 (1)

ORGANIZATION:

ゆうちょ銀行 (2)

金融庁 (1)

Other Examples

- Data mining
- E-Discovery
- Sentiment analysis

Features of Rosette Entity Extractor



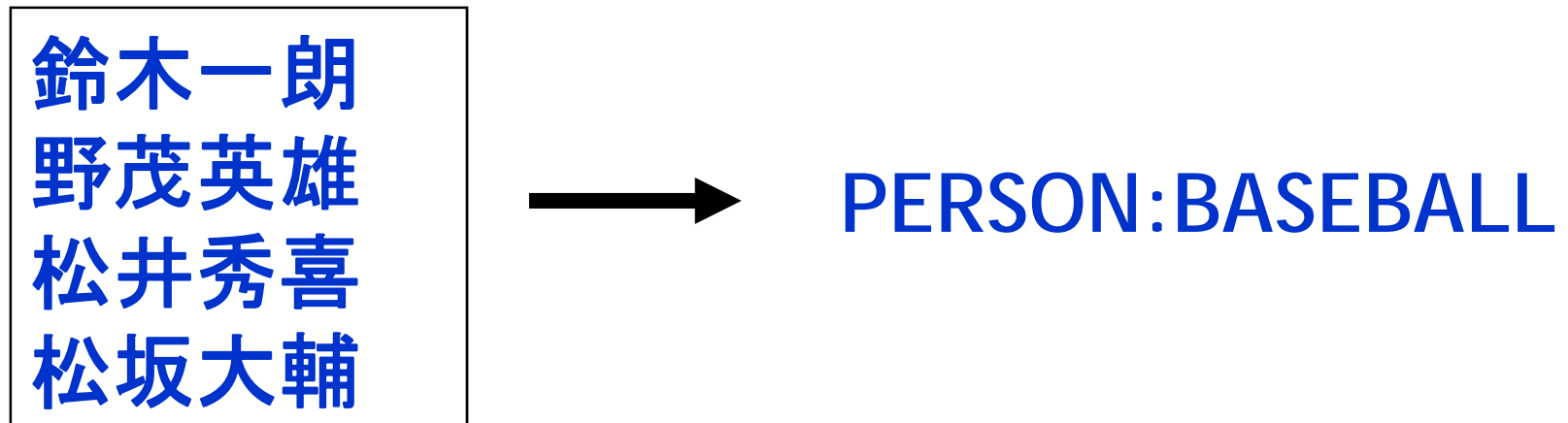
BASIS
TECHNOLOGY

Features of Rosette Entity Extractor

- Gazetteers
- Regular expressions
- Statistical models
- Redactor
- Multilingual

Gazetteers

- Customizable list of words and corresponding entity type



Regular Expressions

- Customizable character pattern and corresponding entity type

[日月火水木金土]曜日



TEMPORAL:DAY_OF_WEEK

Statistical Models

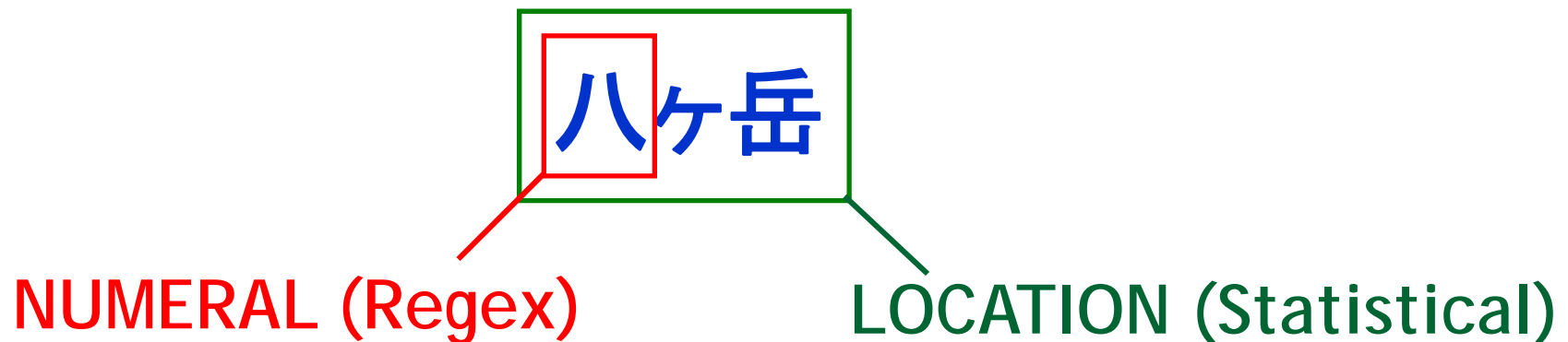
- Context-Sensitive
- Trained on large, manually tagged corpora

川崎さんはさっき川崎を出ました。

PERSON LOCATION

Redactor

- Resolves conflicts between the three entity extractors
- Customizable heuristics, black-lists, precedence



Entities Covered Out-of-the-Box

- Person
- Organization
- Location
- Title
- Religion
- Nationality
- Geopolitical Entity
- Facility

Multilingual

	Person	Organization	Location	Geopolitical Entity	Facility	Religion	Nationality	Title	Credit Card Number	Distance	Email	Latitude / Longitude	Money	Number	Personal ID Number	Phone Number	URL	UTM	Date	Time	
	Statistical / Gazetteer								Regular Expression												
Japanese*	S	S/G	S	--	--	G	G	S	R	R	R	R	R	R	R	R	R	R	R	R	R
Chinese* (SC & TC)	S	S	S	--	--	G	G	S	R	R	R	R	R	R	R	R	R	R	R	R	R
Korean	S	S	S	S	S	G	G	S	R	R	R	R	--	R	R	R	R	R	R	R	R
Arabic*	S	S	S	--	--	G	G	S	R	R	R	R	R	R	R	R	R	R	R	R	R
Persian* (Dari & W. Farsi)	S	S	S	--	--	--	--	G	R	R	R	R	R	R	R	R	R	R	R	R	R
Push <u>to</u> *	S	S	S	--	--	--	--	S	R	R	R	R	R	R	R	R	R	R	R	R	R
Urdu	S	S	S	S	S	S/G	S/G	G	R	--	R	--	R	--	R	R	R	R	--	--	--
English*	S	S	S	--	--	G	G	S	R	R	R	R	R	R	R	R	R	R	R	R	R
English (Uppercase)	S	S	S	--	--	G	G	S	R	R	R	R	R	R	R	R	R	R	R	R	R
French	S	S	S	--	--	--	--	G	R	R	R	R	R	R	R	R	R	R	R	R	R
Italian	S	S	S	--	--	--	--	G	R	R	R	R	R	R	R	R	R	R	R	R	R
German	S	S	S	--	--	--	--	G	R	R	R	R	R	R	R	R	R	R	R	R	R
Spanish	S	S	S	--	--	--	--	S/G	R	R	R	R	R	R	R	R	R	R	R	R	R
Dutch	S	S	S	--	--	--	--	G	R	R	R	R	R	R	R	R	R	R	R	R	R
Portuguese	--	--	--	--	--	--	--	G	R	R	R	R	R	R	R	R	R	R	R	R	R
Hungarian	--	--	--	--	--	--	--	--	R	--	R	R	R	--	R	R	R	R	R	--	--
Czech	--	--	--	--	--	--	--	--	R	--	R	R	R	--	R	R	R	R	R	--	--
Greek	--	--	--	--	--	--	--	--	R	--	R	R	R	--	R	R	R	R	R	--	--
Russian*	S	S	S	--	--	G	G	S/G	R	R	R	R	R	R	R	R	R	R	R	R	R
Polish	--	--	--	--	--	--	--	--	R	--	R	R	R	--	R	R	R	R	R	--	--

- S 2009 Statistical model
- S Statistical model
- G Gazetteer entry
- R Regular expression rule

Thank You!

Benjamin Douglas, Principal Engineer

www.basistech.com





2010 TOKYO

次世代サーチ
テクノロジー
フォーラム

国際文化会館

2010年4月22日
(木)