From Human Language to Useful Information

Gilbane Boston
November 2007

BASIS TECHNOLOGY

Steve Cohen
Founder and EVP
www.basistech.com

Matt Kodama
Product Management
www.endeca.com

About Basis Technology

- Diversified firm specializing in language technology, text analytics, and software internationalization

- Twelve year track record delivering technology to and growing businesses overseas, including: Amazon.com, Ask.com, Google, Kenan Systems, L.L. Bean, Lycos, Ofoto, Overture, PeopleSoft, Progress Software, RightNow

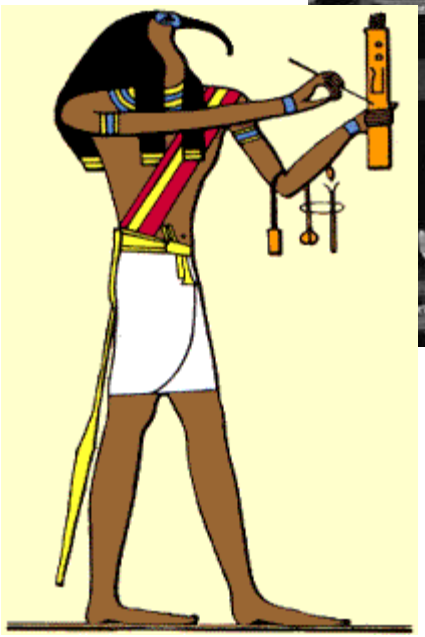
Selected Basis Technology Customers

Information retrieval

- Finding the information you need
  - ⇨ *Structured*
    - Facts
    - Databases, Spreadsheets
    - Single field, row, join, etc. represents information of interest
    - Techniques for retrieval are well developed
  - ⇨ *Unstructured*
    - Text documents
    - Images
    - Audio, video, etc.
    - Also about facts, but these are harder to discern
    - Techniques for retrieval are less developed

Unstructured information retrieval

- Unstructured information retrieval begins and ends with human language



Human language is imprecise

- Choice of language
  - ⇨ *Esperanto, anyone?*
- Choice of writing quality
  - ⇨ *"Formal" news text*
  - ⇨ *"Informal" IM text*
    - CN U RD THIS?
- Choice of spelling
  - ⇨ *Mohammed*
  - ⇨ *Muhamed*
  - ⇨ *Mahmoud*

The New York Times

Techno

WORLD | U.S. | N.Y. / REGION | BUSINESS | TECHNOLOGY | SCIENCE | HEA

Search Tech News & 8,000+ Products

Go

Browse Produ

-- Select a Product

Adding Math to List of Security Threats

By JOHN MARKOFF
Published: November 17, 2007

SAN FRANCISCO, Nov. 16 — One of the world's most prominent cryptographers issued a warning on Friday about a hypothetical incident in which a math error in a widely used computing chip pla
the security of the global electronic commerce system at risk.

Technology to the rescue!

- Machine translation?
- Term translation
- Fuzzy matching
  - ⇨ *Large lists…*
  - ⇨ *Soundex, Phonix*
  - ⇨ *Edit Distance*
  - ⇨ *Machine learned models*
- Name finding

Machine Translation

Now I am alone.
O, what a rogue and peasant slave am I! (555)
Is it not monstrous that this player here,
But in a fiction, in a dream of passion,
Could force his soul so to his own conceit
That from her working all his visage wann'd,
Tears in his eyes, distraction in's aspect, (560)
A broken voice, and his whole function suiting
With forms to his conceit? and all for nothing!
For Hecuba!

English
|
Chinese
|
English

Automatically translated text:

Now, I alone.
O, what a rogue and peasant slave me!
This is not monster, this player here,
However, in novels, in a dream of passion,
Could force his soul, and make our own conceited
This from her work, all his face wann'd,
Tears in his eyes, a part of the distraction,
A broken voice, and his overall functional adaptation
With forms to his conceited? And all why!
For hecuba!

Term translation

- “Taxi” in Chinese
  - ⇨ *出租车*
  - ⇨ *计程车*
  - ⇨ *的士*

Fuzzy Matching

Najaf

Al-Najaf

Annajaf

Muhafaza Najaf

Muhafazat El Najaaf

Muhafazah Annajaf

Muhafaza Al Najaf

Muhafazah Alnajaaf

Muhaafazat Najaf

محافظة النجف

Cross-Lingual Annotation

- Maintain original text
- Add information from target language
- Add linguistic meta-data
  - ⇨ *Entities (people, places, locations)*
  - ⇨ *Part of speech*
  - ⇨ *Word sense*

Rosette Text Analyzer

Dismiss

Results

Click on Arabic Words For Parse Data:

جددت الولايات المتحدة دعوتها للرئيس الباكستاني برويز مشرف لرفع حالة الطوارئ المفروضة على بلادة منذ مطلع الشهر الجاري والعودة إلى النهج الدستوري.

al-Wilayat al-Muttahidah Barwiz Musharraf al-Bakistani

Vocalized	Gloss	Part of Speech	Transliteration
أَلْوِلاَيَات	(the) states/provinces ([fem.pl.])	DET NOUN FEM_PL	al-wilaayaat
أَلْوِلاَيَات	(the) States ([fem.pl.])	DET NOUN FEM_PL	al-wilaayaat

Named Entities

LOCATION

الولايات المتحدة
al-Wilayat al-Muttahidah

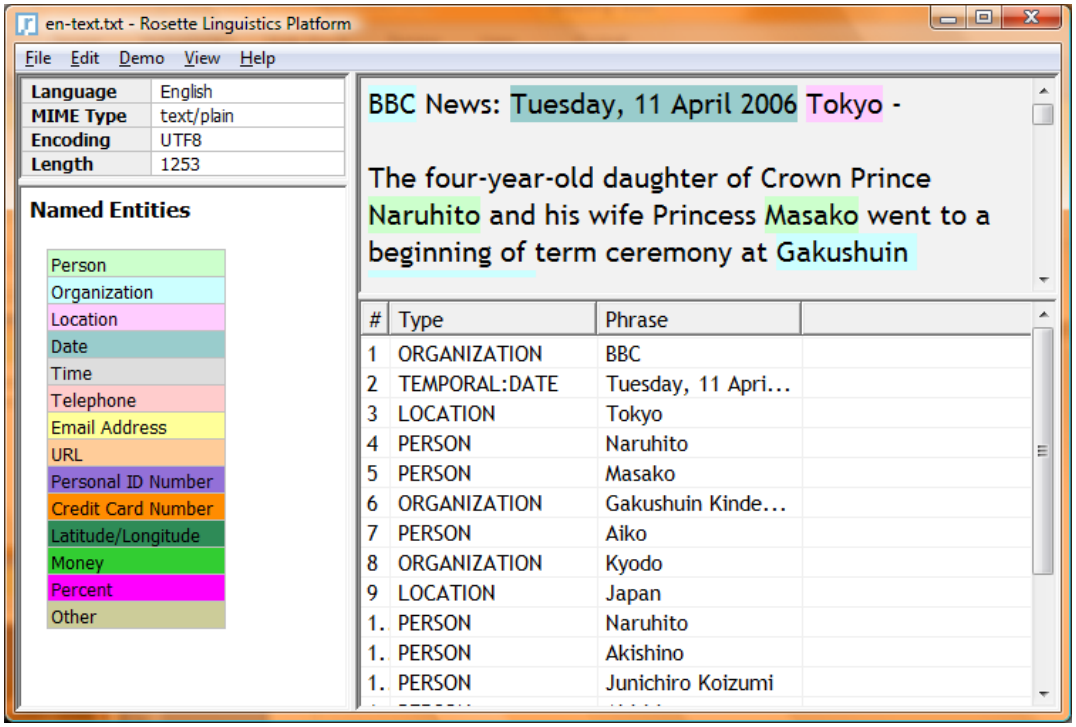
الباكستاني
al-Bakistani

PERSON

برويز مشرف
Barwiz Musharraf

Name Finding

- Find references to people, places, organizations
- Use this “semantic” knowledge about the text to enable retrieval or advanced analysis



Putting it together

"Olmert"

Data Mining and Information Access

Matt Kodama
Product Management

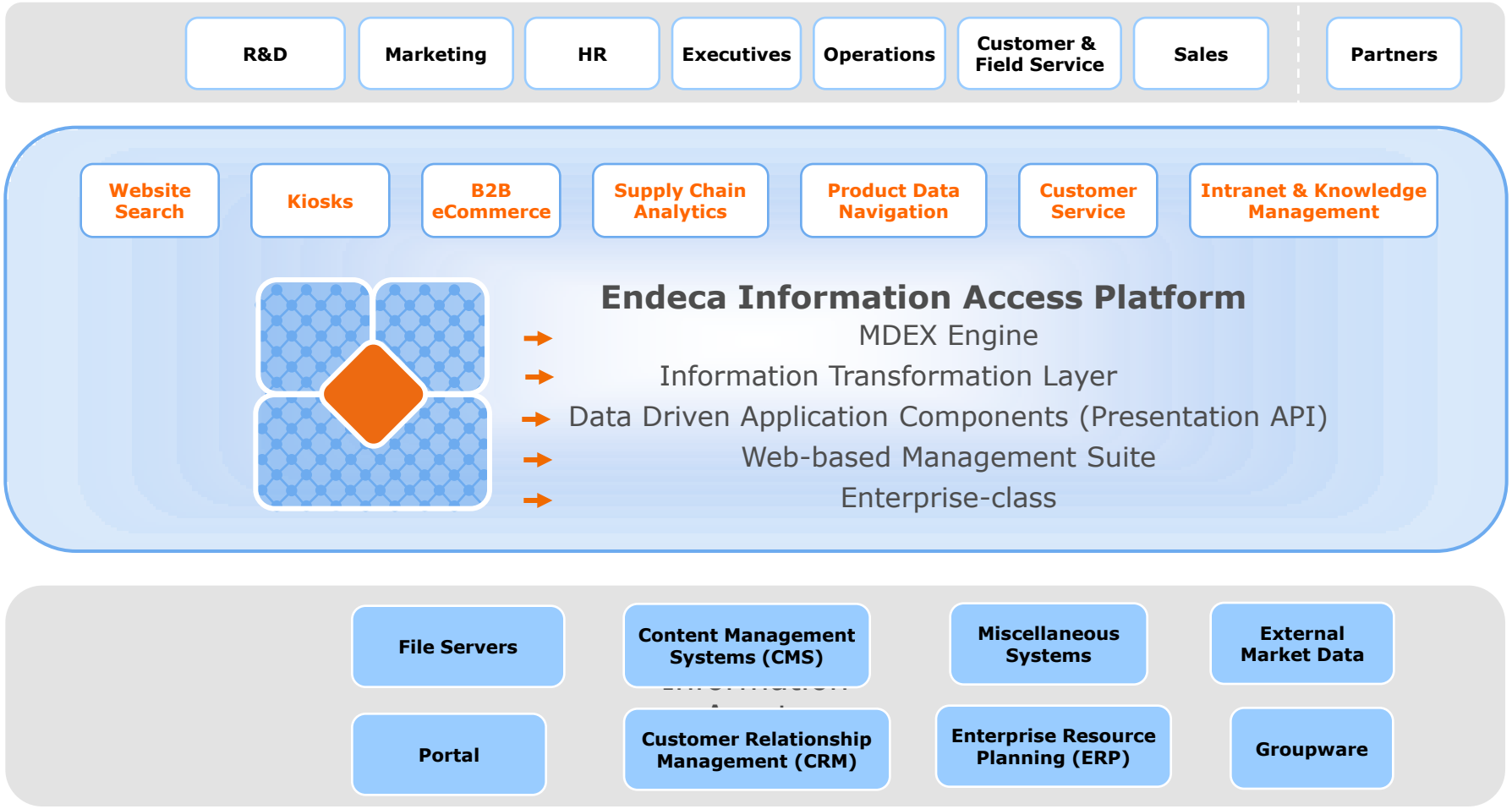
Selected Endeca Customers

Endeca Delivers Business Results

Sample Customers		*Representative Metrics*
AMERICAN EXPRESS	Endeca for Intranet	Content downloads increased 400%
TIBCO The Power of Now®	Endeca for Knowledge Management	Intranet usage increased 30%, call center volume decreased 20%
IBM	Endeca for Human Capital Management	Intranet usage increased 175%
bhpbilliton	Endeca for Knowledge Management	Saved $500M through better consultant staffing analysis
ABN·AMRO	Endeca for Intranet	Significant employee efficiency and productivity improvements

"The solution influences decisions that generate $100's of millions in revenue."

Adaptivity closes the Information Access Gap



Endeca's unique architecture: Adaptivity

Adaptivity is
the dynamic summarization of
erratic data and content
in the current view

Case Study: Arabic in the Intel Community

Search: Within current results Dim Search Text Search In Field: LocOrgPerson GO

Filter By

person_org_loc

- Translated LOCATION [refine]
  - *Ash Sharq Al Awsaṭ (11)*
  - *Al `Irāq (11)*
  - *Baghdād (9)*
  - *Al Amrīkīyah (8)*
  - *`Irāq (8)*
  - *More...*
- Translated ORGANIZATION [ref
  - *al-Hukumah al-'Iraqiyyah (6)*
  - *al-Jaysh al-Amriki (5)*
  - *al-Tayyar al-Sadri (4)*
  - *Tanzim al-Qa'idah (3)*
  - *Hizballah (3)*
  - *More...*
- Translated PERSON [refine]
  - *al-Malikay (7)*
  - *Nuri al-Malikay (7)*
  - *Muqtada al-Sadr (3)*
  - *Jurj Bush (2)*
  - *Bush (2)*
  - *More...*

Breadcrumbs

[x]**Text Search:** All: المالكي

[x]**Clear All Filters**

Record Table | Record Detail

Results 1-10 of 11 **Results per Page:** 10 | 20 | 30 **Page:** 1 2 >

		Document Title
		[x]
1	View Record	العراق يتبرع بـ 35 مليون دولار إلى لبنان
2	View Record	مسؤول عسكري أمريكي: الزرقاوي باقٍ
3	View Record	الصحف العربية: عراق "الجثث" وحكومة "حماس"
4	View Record	القاعدة تؤكد مقتل الزرقاوي وتتوعد بمواصلة الهجمات
5	View Record	بغداد: العثور على 547 جثة مجهولة منذ بداية مايو
6	View Record	صحف:غيتس يهدد بالبقاء طويلا بالمنطقة وحاملة طائرات لمراقبة إيران
7	View Record	إيران تندد باستراتيجية بوش الجديدة بالعراق
8	View Record	مقتل أربعة جنود أمريكيين في الأنبار
9	View Record	الزعيم الشيعي مقتدى الصدر يظهر لأول مرة منذ شهور بالكوفة
10	View Record	بريطانيا تعلن اختطاف خمسة من مواطنيها في العراق

Results 1-10 of 11 **Results per Page:** 10 | 20 | 30 **Page:** 1 2 >

Location Cloud | Person Cloud | Organization Cloud | Tranlsated Cloud

al-Malikay

Nuri al-Malikay

Muqtada al-Sadr Jurj Bush Bush Saddam Husayn al-Zarqawi Abaw Mus'ab al-Zarqawi Mahmud 'Abbas 'Abbas Fu'ad al-Sinyurah Isma'il Haniyyah Byrts Usamah Bin-Ladin Allah Zarqawi 'Abdallah Musa 'Amru Musa Ulmirt Nasrallah al-Asad Shar al-Asad al-Hariri 'Anan More...

Arabic in the Intelligence Community

- Arabic source data
- Analysts not fluent
- Limited access to translators and native speakers
- What to translate?

Prioritize using key terms: person, place, and organization

Data Mining: Entity Extraction

على حافة حرب

Lubnān Al 'Irāq
لبنان يتبرع بـ 35 مليون دولار إلى العراق

1802 (GMT+04:00) - 22/08/06

Baghdād Al 'Irāq
رغم أزمتها الإنسانية -- (CNN) العراق ، بغداد

al-Hukumah al-'Iraqiyyah
الحالية، تبرعت الحكومة العراقية بـ 35 مليون دولار

Lubnān
كمساعدات لهيئة الإغاثة العليا بلبنان ، في إشارة على

al-Hukumah al-Lubnaniyyah
دعم الحكومة اللبنانية ، وتعبير عن مخاوف

Al 'Irāq
العراق من اتساع حجم الأزمة الإنسانية التي يعاني

منها المواطنين اللبنانيين.

Salih Al 'Irāqī
وأعلن التبرع نائب رئيس الوزراء العراقي ، برهم صالح ، حيث قال إنه اتصل هاتفيا لرئيس

Fu'ad al-Sinyurah
الوزراء اللبناني، فؤاد السنيورة ، الاثنين ليبلغه بالتبرعات المالية.

Salih
وذكر مسؤول بارز في مكتب صالح : " لا يمكن أن ندع أشقائنا اللبنانيين يعانون بدون مساعدة

Al 'Irāqī
من أشقائهم العرب. لقد طالب رئيس الوزراء (العراقي) ونائبه، المجتمع الدولي بضرورة بذل

Lubnān
IRIN المزيد من الجهود لوقف العنف الذي انتشر في لبنان ،" وذلك حسب ما نقلته وكالة

Umam al-Muttahidah
بالأمم المتحدة التابعة لخدمات أنباء ومعلومات عمليات الإغاثة الإنسانية.

Lubnān
دمار في لبنان

Persons

- ⇨ *Fuad al-Sinyurah*
- ⇨ *Salih*

Organizations

- ⇨ *Al-Hukumah al-Iraqiyyah*

Locations

- ⇨ *Al Iraq*
- ⇨ *Baghdad*
- ⇨ *Lubnan*

Data Mining: Name Similarity Indexing

- Linguistic analysis of name similarity
  - ⇨ *High likelihood:* *"Al-Malikay" & "Nuri al-Malikay"*
  - ⇨ *Possible:* *"Mahmud Abbas" & "Abbas"*
  - ⇨ *Unlikely:* *"Fu'ad al-Sinyurah" & "Amru Musa"*

- Enables fuzzy expansion of name searches
  - ⇨ *Type "maliki" and search for*
    - → المالكي
    - → al-Malikay
    - → Nuri al-Malikay

Information Access: Fuzzy Keyword Search

Search: khadaffi

Breadcrumbs

[x]**Text Search:** All: القذافي OR قذافي

[x]**Clear All Filters**

- Keyword search
- Stemming / inflection
- Thesaurus
- Spelling correction
- Wildcard
- Phrase interpretation
- DYM suggestions
- …

Information Access: Navigation and Visualization

Filter By

person_org_loc

- Translated LOCATION [refine]
  - *Ash Sharq Al Awsaṭ* (349)
  - *Al Amrīkīyah* (173)
  - *Al Amrīkī* (159)
  - *Al Wilāyāt Al Muttaḥidah* (158)
  - *Wāshinṭun* (118)
  - *More...*
- Translated ORGANIZATION [ret
  - *Tanzim al-Qa'idah* (74)
  - *al-Umam al-Muttahidah* (61)
  - *al-Qa'idah* (59)
  - *al-Barlaman* (37)
  - *Hizballah* (33)
  - *More...*
- Translated PERSON [refine]
  - *Jurj Bush* (78)
  - *Bush* (64)
  - *Usamah Bin-Ladin* (36)
  - *al-Zarqawi* (32)
  - *Mu'ammar al-Qadhafi* (17)
  - *More...*

Location Cloud | Person Cloud | Organization Cloud | Tranlsated Cloud

Jurj Bush
Bush
Usamah Bin-Ladin
al-Zarqawi Saddam
Rayis Mu'ammar al-Qadhafi
al-Qadhafi Zarqawi
Kunduliza Rayis Ayman al-Zawahiri
Musa Abumus'ab al-Zarqawi
'Abdallah al-Thani 'Arafat al-Zarqawi Li Ibn-Ladin
al-Zarqawi Ra's Salih 'Ali Salih al-Nabhan
Nur-al-Din Muhammad Twb 'Abdallah Muhammad Salah
'Abd-al-'Aziz 'Udih Qadhafi Abubakr JAnjlAny
Muhammad 'Ali Hamadi Rubirt Stythm
Jamal Ahmad Muhammad 'Ali al-Badawi More...

Information Access: Transliteration in Context

Record Table | Record Detail

Results 1-10 of 11 | Results per Page: 10 | 20 | 30 | Page: 1 2 >

		Document Title
		[x]
1	View Record	العراق يتبرع بـ 35 مليون دولار إلى لبنان
2	View Record	مسؤول عسكري أمريكي: الزرقاوي بالتأكيد
3	View Record	الصحف العربية: عراق "الجثث" وحكومة "حماس""
4	View Record	القاعدة تؤكد مقتل الزرقاوي وتتوعد بمواصلة الهجمات
5	View Record	بغداد: العثور على 547 جثة مجهولة منذ بداية مايو
6	View Record	صحف:غيتس يهدد بالبقاء طويلا بالمنطقة وحاملة طائرات لمراقبة إيران
7	View Record	إيران تندد باستراتيجية بوش الجديدة بالعراق
8	View Record	مقتل أربعة جنود أمريكيين في الأنبار
9	View Record	الزعيم الشيعي مقتدى الصدر يظهر لأول مرة منذ شهور بالكوفة
10	View Record	بريطانيا تعلن اختطاف خمسة من مواطنيها في العراق

Results 1-10 of 11 | Results per Page: 10 | 20 | 30 | Page: 1 2 >

على حافة حرب

Lubnān لبنان يتبرع بـ 35 مليون دولار إلى Al `Irāq العراق

1802 (GMT+04:00) - 22/08/06

Lubnān دمار في لبنان

رغم أزمتها الإنسانية -- (CNN) Al `Irāq العراق ، Baghdād بغداد

الحالية، تبرعت al-Hukumah al-'Iraqiyyah الحكومة العراقية بـ 35 مليون دولار

كمساعدات لهيئة الإغاثة العليا Lubnān بلبنان ، في إشارة على

دعم al-Hukumah al-Lubnaniyyah الحكومة اللبنانية ، وتعبير عن مخاوف

Al `Irāq العراق من اتساع حجم الأزمة الإنسانية التي يعاني

منها المواطنين اللبنانيين.

وأعلن التبرع نائب رئيس الوزراء Al `Irāqī العراقي ، برهم Salih صالح ، حيث قال إنه اتصل هاتفيا لرئيس

الوزراء اللبناني، Fu'ad al-Sinyurah فؤاد السنيورة ، الاثنين ليبلغه بالتبرعات المالية.

وذكر مسؤول بارز في مكتب Salih صالح : " لا يمكن أن ندع أشقائنا اللبنانيين يعانون بدون مساعدة

من أشقائهم العرب. لقد طالب رئيس الوزراء Al `Irāqī (العراقي) ونائبه، المجتمع الدولي بضرورة بذل

IRIN المزيد من الجهود لوقف العنف الذي انتشر في Lubnān لبنان ،" وذلك حسب ما نقلته وكالة

Umam al-Muttahidah بالأمم المتحدة التابعة لخدمات أنباء ومعلومات عمليات الإغاثة الإنسانية.