Rosetta: An Analyst’s Co-Pilot

Salim Roukos
IBM TJ Watson Research Center
OUTLINE

- Overview of GALE tasks
- Analysis of HTER GALE results
- Speech-To-Text overview
- Direct Translation Model II
- UIMA: Interoperability
- TALES demo
GALE

Language Exploitation Environment

Processing Engines

Transcription

Translation

Distillation

Interaction

Military Command or personnel

English-speaking analyst
GALE Rosetta Team

ROSETTA TASKS and TEAM

Transcription
- IBM
- Brown
- CMU
- JHU

Translation
- IBM
- CMU
- JHU
- Stanford
- UMD

Distillation
- IBM
- CMU
- Pitt
- UMD
- Stanford

Focus on A, C, E
Goals for ROSETTA System

- Ingest traditional and informal media:
  - broadcast news, talk shows, …
  - Newswire, news web sites, blogs, …

- Scale to large volumes of multimodal/multilingual inputs
  - Accurate, robust, quickly deployable engines, near real-time (up to 3x), 24x7, …

- Start w/Arabic, Chinese, English; scalable to 10’s of languages

- Adaptive to user needs -- Personalized digests
  - Robust, explainable, and controllable models of user and task
  - Automatic generation of focused reports & graphics, …

- End2End system as living laboratory
  - Continuous testing
ROSETTA TASKS: LEE

- Accelerate research & speedup insertion

UIMA

- Common Annotation Structure (CAS) as input/output of multimodal processing engines/annotators/components
- Plug&Play: composition/integration of UIMAfied components
- Local/remote components with different OS’s
- Open source

Rosetta will create:

- Common Type System
- Common Repository for componentry

- MEMT: combine multiple MT engines
ROSETTA TASKS (continued):

- **Transcription**
  - Tightly integrated translation: small marginal error rate by combining speech-to-text and translation
  - 3xRT or less runtime: fast, reliable, deployable system using common structure across languages and genres

- **Translation**
  - Preserving meaning: who did what to whom
  - Confidence measures: reducing human correction/editing

- **Distillation**
  - End2End system: task based eval. of improved components
  - Entity/relations networks, adaptive tracking, focused summarization, user modeling
GNG (To Go or Not To Go:-) Evaluation

- Transcription and Translation (HTER)
  - Human post edits system output
    - Editor makes “minimum edits” of system output to reproduce correct meaning
    - HTER: Human Translation Error Rate
    - Control for human instruction across conditions/years – re-use fixed set of error full translations
  - YEAR1: GNG edit distance
    - Transcription: 65% accuracy
    - Translation: 75% accuracy
  - YEAR5: Both at 95%
DISTILLATION Evaluations

- **GO/NOGO**
  - Compare automatic system output to human
  - YEAR1: machine 50% of human using chosen metric

- **UTILITY**
  - Compare human output in a task using either baseline or GALE system
  - Open spec -- showcase technology
DISTILLATION GNG: Sample NL Question Schemata 1

Two types of questions: OPEN and SPECIFIC

OPEN:
- LIST FACTS ABOUT EVENTS DESCRIBED ASfolLOWS: z
- WHAT [people/org/countries] ARE RELATED TO y:event AND HOW?
- PRODUCE A BIOGRAPHY OF [person]
- PROVIDE INFORMATION ON [organization]
- FIND STATEMENTS MADE BY OR ATTRIBUTED TO [person] ON [topic(s)]
- DESCRIBE THE RELATIONSHIP OF [person/org] TO [person/org]
- DESCRIBE [topic(s)] AND INVOLVEMENT OF [country]
- DESCRIBE THE PROSECUTION OF [person] FOR [crime]
- HOW DID x:country REACT TO y:event?
- WHAT CONNECTIONS ARE THERE BETWEEN [event 1/topic 1] and [event 2/topic 2]?
SPECIFIC:

- FIND MUTUAL ACQUAINTANCES OF [person] AND [person]
- TELL ME ABOUT [person’s] MEETINGS ON [topic]
- FIND PASSAGES ABOUT [attacks] BY/OR ATTRIBUTED TO [group]
- FIND PASSAGES ABOUT [attacks] {IN [location] DURING [time interval]}
- DESCRIBE OUTBREAKS OF [disease] {IN [region] IN [time period]}
- IDENTIFY PERSONS ASSOCIATED WITH [organization] WHO HAVE BEEN INDICTED ALONG WITH HOW THEY'RE RELATED
- IDENTIFY PERSONS ARRESTED FROM [organization] AND GIVE THEIR NAME AND ROLE IN ORGANIZATION AND TIME AND LOCATION OF ARREST
- DESCRIBE ATTACKS in [location] DURING THE PAST [duration] GIVING LOCATION (AS SPECIFIC AS POSSIBLE), DATE, AND NUMBER OF DEAD AND INJURED
- WHERE HAS [person] BEEN AND WHEN?
GALE Transcription & Translation GNG Evaluation

- **Arabic and Chinese**
  - Speech
    - Broadcast News (BN) 10kw
    - Broadcast Conversation (BC) 10kw
  - Text
    - Newswire (NW) 10kw
    - NewsGroup/WebLog (WL) 10kw

- **1 Gold Reference with some word/phrase alternations**

- **3 Consortia participated in GALE06 Eval**
  - Agile (BBN)
  - Nightingale (SRI)
  - Rosetta (IBM)
Bleu vs. TER

$R^2 = 0.9039$

ArNW

ZhNW

ArBN

ZhNG

ArNG

ArNW-ASB

ZhBC

ZhBN
BLEU vs. TER

- Agile
- Nightingale
- Rosetta

$R^2 = 0.9039$
$R^2 = 0.8869$
$R^2 = 0.9822$
HTER

- Human editors post-edit MT output to get same meaning as reference translation
- HTER (Human Translation Error Rate)
  - Count all the edit operations

\[ HTER = \frac{I + D + S + M}{|R|} \]

- M is number of word or phrase shift movements
LDC multipass Post Editing

<table>
<thead>
<tr>
<th>Rosetta</th>
<th>P1</th>
<th>P2</th>
<th>FINAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>NW</td>
<td>21.2%</td>
<td>19.8%</td>
<td>16.5%</td>
</tr>
<tr>
<td>Delta</td>
<td>-1.4%</td>
<td>-4.7%</td>
<td></td>
</tr>
<tr>
<td>R2</td>
<td>90%</td>
<td>96%</td>
<td></td>
</tr>
</tbody>
</table>
The French President to Visit India to Intensify Bilateral Cooperation

New Delhi 16 February (Xinhua) said Naftyj Sarna, spokesman for the Indian Foreign Ministry in New Delhi today, Thursday, that the French President, Jacques Chirac will visit India on 19 and 20 Of February.

It is expected to be the signing of a number of agreements and memoranda of understanding during the visit reflecting the extent of the cooperation between India and France.

Such agreements include a declaration on the development of nuclear energy for peaceful purposes, and on cooperation in the field of defense, and a memorandum of understanding on cooperation in the field of tourism.

The two countries aim to intensify bilateral cooperation in various fields, including their partnership in the political, economic, defense, space, and civilian nuclear energy.

President Jacques Chirac will deliver a keynote speech on economic partnership between India and France.

President Chirac is accompanied in the visit by his wife Bernadette Chirac, and the ministers of foreign affairs, defense, economy, finance, industry, foreign trade, tourism as well as some 30 senior managers of major French companies.
The Economic Offer: for Environment-friendly Cars in the Chinese Market/First and Last Addition/ HTER=0%
He pointed out that the two official tests on the Al-Hajeen, which indicates the start of mass production of environment-friendly in China. HTER=26%
He added a senior official of the Ministry of Science and Technology that China has achieved remarkable progress in developing the cars will increase local production without doubt their competitiveness in the global market. HTER=15%

The Economic Offer: for Environment-friendly Cars in the Chinese Market/First and Last Addition/
Wan pointed out that the two hybrid bus types passed official tests, which indicates the start of mass production of environment-friendly buses in China.
A senior official of the Ministry of Science and Technology added that China has achieved remarkable progress in developing the cars and local production without doubt will increase their competitiveness in the global market.
Can we predict document HTER from document BLEU/TER?

Doc BLEU= 0.25 => Doc HTER= 16.5% +/- SE

<table>
<thead>
<tr>
<th>NW TEXT</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>STD. ERR.</td>
<td>TER</td>
<td>BLEU</td>
</tr>
<tr>
<td>Doc=302wd</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Agile</td>
<td>5.0</td>
<td>5.7</td>
</tr>
<tr>
<td>Nightingale</td>
<td>5.8</td>
<td>5.7</td>
</tr>
<tr>
<td>Rosetta</td>
<td>5.3</td>
<td>5.5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>BN AUDIO</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>STD. ERR.</td>
<td>TER</td>
<td>BLEU</td>
</tr>
<tr>
<td>Doc=770wd</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Agile</td>
<td>4.5</td>
<td>4.9</td>
</tr>
<tr>
<td>Nightingale</td>
<td>6.6</td>
<td>4.5</td>
</tr>
<tr>
<td>Rosetta</td>
<td>4.2</td>
<td>4.5</td>
</tr>
</tbody>
</table>

To be 95% confident of passing a GNG threshold one needs
100 docs (for a stderr of 0.5% in HTER) around that level:

=> need DEV SETS of 1000 docs per condition
Can we predict document HTER from document Post Editing @IBM?

Subset of Arabic NW: 18 docs Post-Edited @ IBM

<table>
<thead>
<tr>
<th>Post Editing</th>
<th>Agile</th>
<th>Nightingale</th>
<th>Rosetta</th>
</tr>
</thead>
<tbody>
<tr>
<td>LDC HTER</td>
<td>21.01%</td>
<td>20.18%</td>
<td>19.19%</td>
</tr>
<tr>
<td>IBM HTER</td>
<td>34.02%</td>
<td>32.94%</td>
<td>32.91%</td>
</tr>
<tr>
<td>R2</td>
<td>62%</td>
<td>59%</td>
<td>58%</td>
</tr>
<tr>
<td>STD ERR</td>
<td>5.9%</td>
<td>5.0%</td>
<td>5.9%</td>
</tr>
</tbody>
</table>

- Similar results for Chinese
The 2006 Rosetta Transcription Effort
Net Rosetta Progress This Year

<table>
<thead>
<tr>
<th></th>
<th>Mandarin (RT04 Test set)</th>
<th>Arabic (RT04 Test set)</th>
</tr>
</thead>
<tbody>
<tr>
<td>December</td>
<td>23.2%</td>
<td>21.7%</td>
</tr>
<tr>
<td>June</td>
<td>13.5%</td>
<td>12.6%</td>
</tr>
<tr>
<td>Improvement</td>
<td>42%</td>
<td>42%</td>
</tr>
</tbody>
</table>
Where did the improvement come from?

Arabic

- Algorithmic design
- Unsupervised Data: 750 hrs
- LDC GALE Y1 DATA: 50 hours, AM; 200 hrs. LM

Mandarin

- Algorithmic design
- TDT-4 Lightly supervised Data
- LDC Y1 Data: 450 hours
Transcription Flow Charts

**Arabic:**

1. Segmentation
2. Unvowelized Decoding (SA)
3. Vowelized Decoding (SA++)
4. Adaptive LM Rescoring

- 15.3% wer, bnat
- 13.7%
- 13.4%

* Numbers on subset of BNAT and BCAD

**Mandarin:**

1. Segmentation
2. CMU Self-Adapted Decoding
3. IBM X-Adapted Decoding
4. JHU Rescored Lattice
5. IBM Rescored Lattice
6. CNC

- 18.4% cer
- 14.8%
- 13.9%
- 14.0%
- 13.7%

* Numbers on subset of LDC2006E10 and dev05bcm
What happened between Sep’05 and July’06?

- And the improvements come from …
- LDC data : 1.2%
- Unsupervised Training : 1.3%
- Vowelization : 2.0%
- Big Vocabulary : 1.5%
- Cross-Adaptation Unvowelized-Vowelized : 1.0%
Pronunciation Probabilities

- Vowelized Setup: 617k vocabulary, 2m pronunciations
- Forced alignment on training data (incl. unsupervised BN-03)

<table>
<thead>
<tr>
<th>Pron. Prob.</th>
<th>RT-04</th>
<th>BNAT-05</th>
<th>BCAD-05</th>
</tr>
</thead>
<tbody>
<tr>
<td>no</td>
<td>16.0%</td>
<td>17.3%</td>
<td>26.0%</td>
</tr>
<tr>
<td>yes</td>
<td>14.9%</td>
<td>16.4%</td>
<td>25.1%</td>
</tr>
</tbody>
</table>

- Developed technology to cope with 2 million pronunciations
- Significant improvements from pronunciation probabilities
Vowelization and Broadcast Conversations ..

- ML models: VTLN, FMLLR, MLLR

<table>
<thead>
<tr>
<th></th>
<th>RT-04</th>
<th>BNAT-05</th>
<th>BCAD-05</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unvowelized</td>
<td>17.0%</td>
<td>18.7%</td>
<td>25.4%</td>
</tr>
<tr>
<td>Vowelized</td>
<td>14.9%</td>
<td>16.4%</td>
<td>25.1%</td>
</tr>
</tbody>
</table>

- Significant improvements on Broadcast News, but not on Broadcast Conversations ! -> Need to investigate:
  - Dialect issue?
  - BC training data with vowelized transcripts?
## Evaluation Results

<table>
<thead>
<tr>
<th></th>
<th>BC</th>
<th>BN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabic</td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Dev</td>
<td>21.5</td>
<td>13.7</td>
</tr>
<tr>
<td>- Test</td>
<td>34.0</td>
<td>24.4</td>
</tr>
<tr>
<td>- HTER</td>
<td>35.6</td>
<td>29.2</td>
</tr>
<tr>
<td>Mandarin</td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Dev</td>
<td>20.7</td>
<td>12.9</td>
</tr>
<tr>
<td>- Test</td>
<td>24.1</td>
<td>13.4</td>
</tr>
<tr>
<td>- HTER</td>
<td>37.1</td>
<td>32.4</td>
</tr>
</tbody>
</table>

- Really big mismatch between dev & test
- Some mismatch between dev & test
- We hit the target!
One Key Lesson: Need wider variety of training data

Very little training data for LBC – poor results on test set.
In the future we would like to have at least 10h of speech from each source.
Predicting the WER on New Test Sets
Motivation

- Rapidly assess the performance of an ASR system on a new test set without the need of a reference transcript

- Creating an accurate reference is a time-consuming process
  - Expertise may not be readily available (e.g. foreign languages)
  - Have to rely on other institutions to provide reference (e.g. NIST)

- Applications
  - Predict system performance in government evaluations 😊
  - Select data for (un)supervised training (active learning)
  - Change system configuration to minimize predicted WER
How can we compute $\text{WER}_{A'}$?

Training: all WERs known

Test: only $\text{WER}_{A'B'}$ known
How can we compute \( \text{WER}_{A'} \)?

Training: all WERs known

Test: only \( \text{WER}_{A'B'} \) known
Performance on the 2006 GALE evaluation data
Performance on the 2006 GALE evaluation data
Performance on the 2006 GALE evaluation data

True WER=29.2%, predicted WER=30.0%, CORR=0.87, MAD=5.4
Rosetta:
MT GALE GnG06 Report
A Direct Translation Model II
How many phrases do we need?

- N-M blocks (Used by most SMT systems)
  - General
    - All possible blocks extracted
    - 40-50M blocks in Arabic
    - Sparsity problems
DTM Decoder (aka MaxEnt)

- **Block style**
  - Allow variables in target sequences
  - 1-M blocks
    - Part of a minimalist system
    - Typical size 1.6M blocks

- **Utilizing English, Arabic analysis**
  - Segmentation, POS
  - POS

- **Feature functions on streams of information**

- **Framework for parameter estimation**

```
PREP NN NSUFF_FEM
lljnp → of the VAR committee
Almrkzyp → central
```
Direct Translation Model

- **Joint future: Jump, Target Sequence**
  \[ p(T, j | S) \]
  - \( j \) = jump, which is the number of positions from the previously translated source word position
  - Integrates Distortion and Word-selection model

- **Features**
  - Lexical:
    - Left and Right context of source sequences
    - Questions about the left context of a target sequence
  - Part-of-speech, Segmentation

- **Features shared across phrase blocks**
  - Feature parameters trained to maximize log-likelihood
    - **No direct optimization of any translation quality metric** (BLEU, TER)

- **Details in an upcoming paper**
Features

- **MaxEnt Block Example**
  33 0.0876793 0.0274136 | llHzb | of the VAR_1 party | 0 0 -1 0 || l# l# Hzb

- **Block Internal: Seg Features**

<table>
<thead>
<tr>
<th>Cnt</th>
<th>Alpha</th>
<th>Jump</th>
<th>Tgt</th>
<th>Seg</th>
</tr>
</thead>
<tbody>
<tr>
<td>1107</td>
<td>1.047</td>
<td>-2</td>
<td>of</td>
<td>l#</td>
</tr>
<tr>
<td>3120</td>
<td>0.989</td>
<td>-1</td>
<td>of</td>
<td>l#</td>
</tr>
<tr>
<td>55461</td>
<td>1.319</td>
<td>1</td>
<td>of</td>
<td>l#</td>
</tr>
<tr>
<td>7009</td>
<td>1.225</td>
<td>2</td>
<td>of</td>
<td>l#</td>
</tr>
</tbody>
</table>

- **Block Context Feature**
  - 11 1.66021 0.0330579 1024 -1 party llHzb // communist AI$ywEy chinese AI$Syny

- **New Feature ~ coding time + 8 hours training + 1 hr decode time**
## Experiments - NIST

<table>
<thead>
<tr>
<th>Feature Types</th>
<th># of feats (MT05)</th>
<th>MT-05</th>
<th>MT-06 (NIST)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MaxEnt Decoder</td>
<td>520,210</td>
<td>48.21</td>
<td></td>
</tr>
<tr>
<td>Lexical Feats</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>+Lexical Context</td>
<td>1,551,582</td>
<td>49.24</td>
<td></td>
</tr>
<tr>
<td>+Segmentation Feats</td>
<td>3,063,023</td>
<td>49.51</td>
<td></td>
</tr>
<tr>
<td>+Part-of-Speech Feats</td>
<td>3,370,901</td>
<td>49.87</td>
<td></td>
</tr>
<tr>
<td>+Distortion Feats</td>
<td>3,412,210</td>
<td>49.98</td>
<td>38.61</td>
</tr>
<tr>
<td>Block Decoder</td>
<td></td>
<td>49.06</td>
<td>36.92</td>
</tr>
</tbody>
</table>
UIMA: ARCHITECTURE FOR DARPA GALE

- Highly-distributed plug-and-play architecture
- Support for multi-modal sources
- Support for local/remote heterogenous components
- Open Source
UIMA’s Basic Building Blocks are Annotators. They iterate over an artifact to discover new types based on existing ones and update the Common Analysis Structure (CAS) for upstream processing.
Common Annotation Structure (CAS):
Multiple Subject of Analysis (SOFA) in CAS Supports Multi-Modal Analysis

- Multiple views of an artifact can each support independent sets of attributes
- Focus can changes from audio to text to both
- Attributes directed to one or more SOFAs
A common platform for development, composition and deployment of multi-modal analytics into different carriers.

Development Tools
Component Repository

UIMA Framework

UIMA Analysis Components

Large-Scale Distributed Deployment
Standalone
Product Embeddings

Web Fountain, ...
UIMA SDK
OmniFind, WPS, LWP...

Powered by UIMA-Compliant Run-Times
IBM dives deeper into corporate search

Published: August 7, 2005, 9:01 PM PDT

By Elinor Mills
Staff Writer, CNET News.com

IBM is promoting a new standard to allow interoperability between software that helps corporations search for and analyze unstructured data across their corporate networks, including e-mails, Word documents and anything that is not formatted in columns and rows.

The company was set to release on Monday a new version of its WebSphere Information Integration OmniFind Edition corporate information management tool. It integrates technology called Unstructured Information Management Architecture (UIMA) that IBM designed to improve the processing of text within documents and other unstructured content sources to help find relationships and meaning beyond just keywords.

IBM, a longtime supporter of the open-source movement in which developers freely write and modify software and share code, also is presenting UIMA to the Open Source Technology Group, a network of online technology resources. The updated software tool is available from IBM now and is expected to be available through the SourceForge developers Web site by the end of the year.

*IBM has been investing in a huge initiative since 2001 in information*
Overview | Transcription | Translation | Distillation

Rosetta

Oct 2006

2-minute-segmented Arabic video files

STT1 (BBN)

STT2 (CMU)

STT3 (IBM)

MT1 (LW)

MT2 (IBM)

MT3 (RWTH)

MEMT (CMU)

Extract Audio

Generate Story-boards

Perl

Windows 2003 Server

Java/C++

Windows 2003 Server

Java/C++/Perl

Linux

Java/C++

Linux

C++

Linux

Story-board-annotated video files

Perl

Windows 2003 Server

Java/C++

Windows 2003 Server

Java/C++/Perl

Linux

Java/C++

Linux

C++

Linux

2-minute-segmented Arabic video files

Perl

Windows 2003 Server

Java/C++

Windows 2003 Server

Java/C++/Perl

Linux

Java/C++

Linux

C++

Linux

Story-board-annotated video files
ProDasa Flow: Serial

Legend
- Remote Engines

Views of Data
- Audio
- Text

- 2-minute-segmented Arabic video files
- Extract Audio
- STT1 (BBN)
- STT2 (CMU)
- STT3 (IBM)
- Determine Sentence Units
- MT1 (LW)
- MT2 (IBM)
- MT3 (RWTH)
- MEMT (CMU)
- Generate Storyboards

Audio

NIGHTINGALE
Rosetta
IOD Enables On-Line MEMT, Increased Accuracy

- GNG Arabic speech test set (34 of 37 audio files)
- Case-insensitive evaluation

<table>
<thead>
<tr>
<th>System</th>
<th>TER</th>
<th>BLEU4</th>
<th>BLEU1</th>
<th>METEOR</th>
</tr>
</thead>
<tbody>
<tr>
<td>STT A, MT Y</td>
<td>75.9</td>
<td>0.100</td>
<td>0.349</td>
<td>0.405</td>
</tr>
<tr>
<td>STT A, MT Z</td>
<td>75.4</td>
<td>0.097</td>
<td>0.366</td>
<td>0.396</td>
</tr>
<tr>
<td>STT B, MT Y</td>
<td>74.7</td>
<td>0.101</td>
<td>0.340</td>
<td>0.405</td>
</tr>
<tr>
<td>STT B, MT Z</td>
<td>74.7</td>
<td>0.094</td>
<td>0.334</td>
<td>0.395</td>
</tr>
<tr>
<td>MEMT</td>
<td>75.7</td>
<td>0.116</td>
<td>0.421</td>
<td>0.440</td>
</tr>
<tr>
<td>MEMT % gain</td>
<td>-1</td>
<td>+15</td>
<td>+15</td>
<td>+9</td>
</tr>
</tbody>
</table>
GNG Results vs. IOD

- Research systems ~50% better than product engines
- Case-sensitive GNG vs. case-insensitive IOD
- Significant work to productize
TALES: Multimodal Trans-lingual Analytics

- Internet
  - HTML, PDF, ...
  - JPG, BMP, ...
- Satellite Broadcast
  - Arabic TV
  - Chinese TV

Real-time Analysis Engines

Search Indexes

Data Store

Search Engines

Query Translation

Analyst Applications

Data available as quickly as acquired
- 5 min delay on video content
- 15 min delay on web pages

- Speech-to-text
- Statistical machine translation
- Cross-lingual search
Video Processing Flow

4 TV Channels:
- Al-Jazeera
- CCTV4
- etc.

Key Frame Extraction → Speech/Non-Speech Detection → Speaker Segmentation & Clustering → Speech-To-Text Arabic Chinese → Name-Entity Detection → Machine Translation Arabic-English Chinese-English → Name Entity Translation & Transliteration

Token Models → N-E Models → Trans Models

Content Manager → Search Indexes → Storage Filter → Indexer → Model Adapt

OmniFind
Text Processing Flow

1. **Duplicate Detection**
2. **Language ID**
3. **Detag**
4. **Tokenize Text**
5. **Name-Entity Detection**
6. **Machine Translation**
7. **Name Entity Translation & Transliteration**

**UIMA Analytics**

- **Content Manager**
- **Storage Filter**
- **Indexer**
- **Model Adapt**

**OmniFind**
TALES Foreign Broadcast Video Monitoring and Search System

- UIMA-based trans-lingual search technology:
  - Speech-to-Text
  - Machine Translation (English, Arabic, Chinese)
  - Advanced Text Analysis (language identification and translation, named entity extraction and translation)
  - Cross-lingual Information Retrieval
Thankyou

謝謝

شكرًا