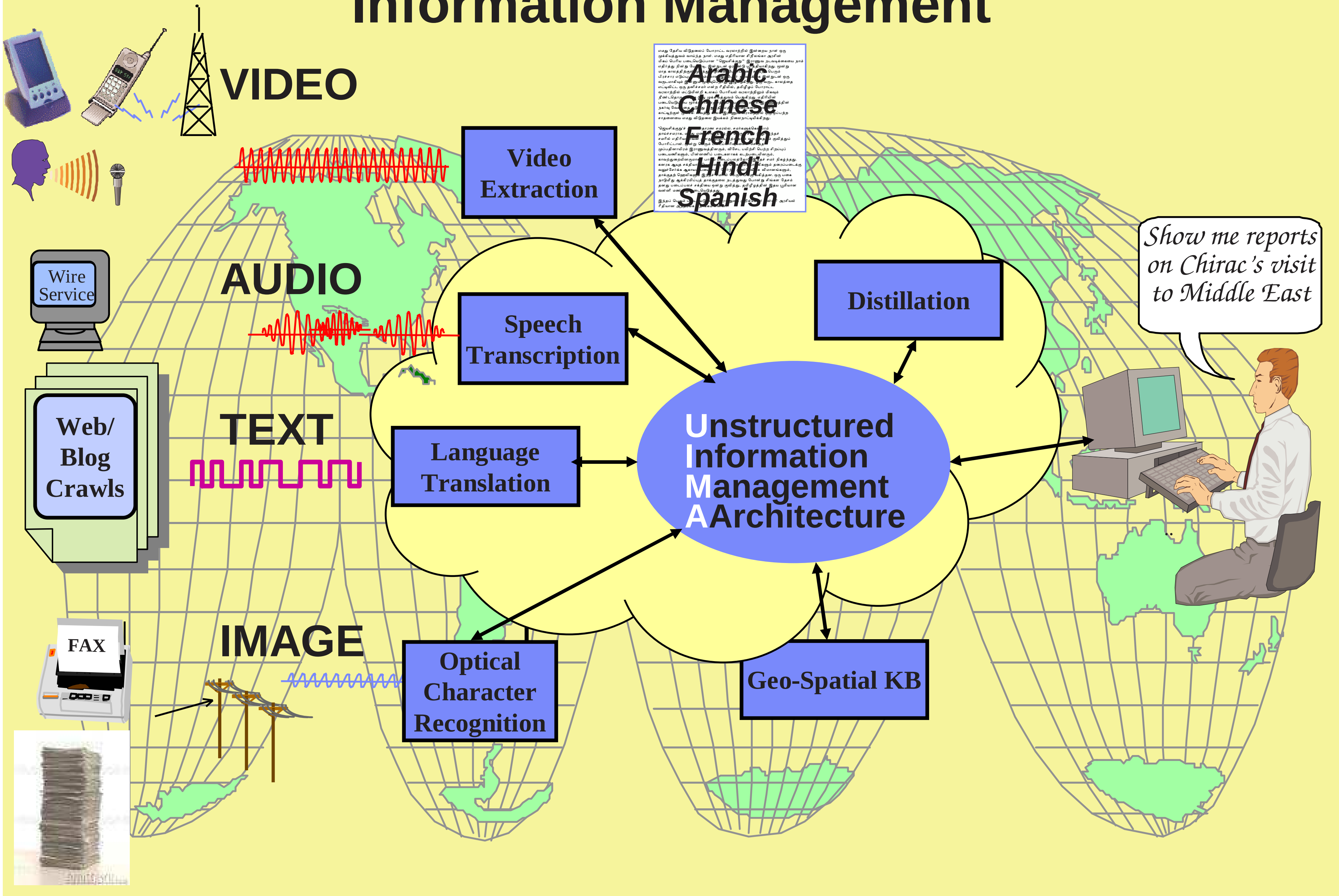
Rosetta:
An Analyst's Co-Pilot

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

Multi-Lingual & Multi-Modal Information Management



VIDEO

AUDIO

TEXT

IMAGE

Wire Service

Web/ Blog Crawls

FAX

Arabic
Chinese
French
Hindi
Spanish

Video Extraction

Speech Transcription

Language Translation

Optical Character Recognition

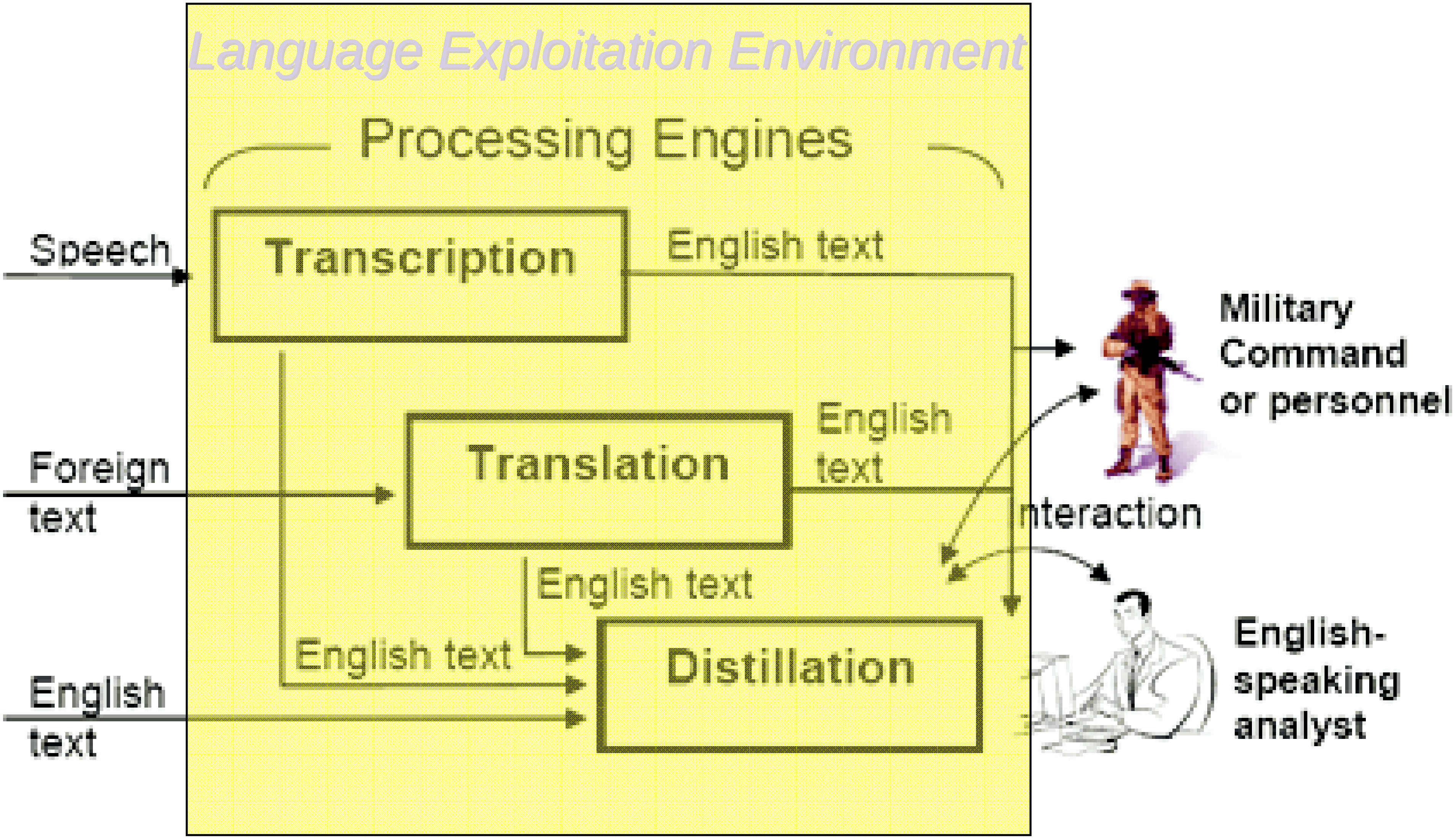
Distillation

Geo-Spatial KB

Unstructured
Information
Management
AArchitecture

Show me reports on Chirac's visit to Middle East

GALE



Language Exploitation Environment

Processing Engines

Speech

Transcription

English text

Foreign text

Translation

English text

English text

English text

Distillation

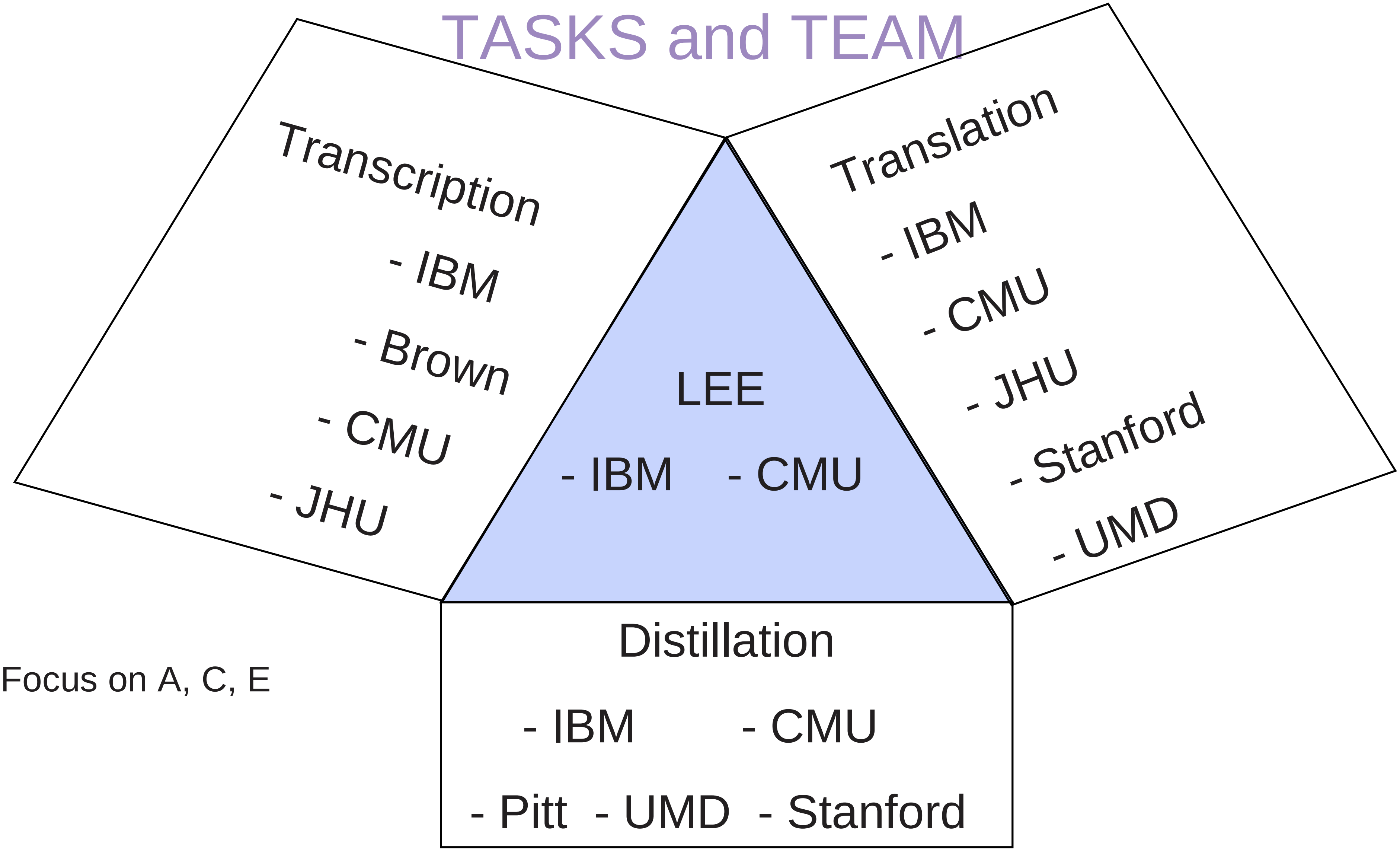
English text

Military Command or personnel

nteraction

English-speaking analyst

ROSETTA
TASKS and TEAM



Transcription

- IBM
- Brown
- CMU
- JHU

Translation

- IBM
- CMU
- JHU
- Stanford
- UMD

LEE

- IBM - CMU

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 usin 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 component**
  - **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 I

Two types of questions: OPEN and SPECIFIC

OPEN:

- **LIST FACTS ABOUT EVENTS DESCRIBED AS FOLLOWS: 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]?**

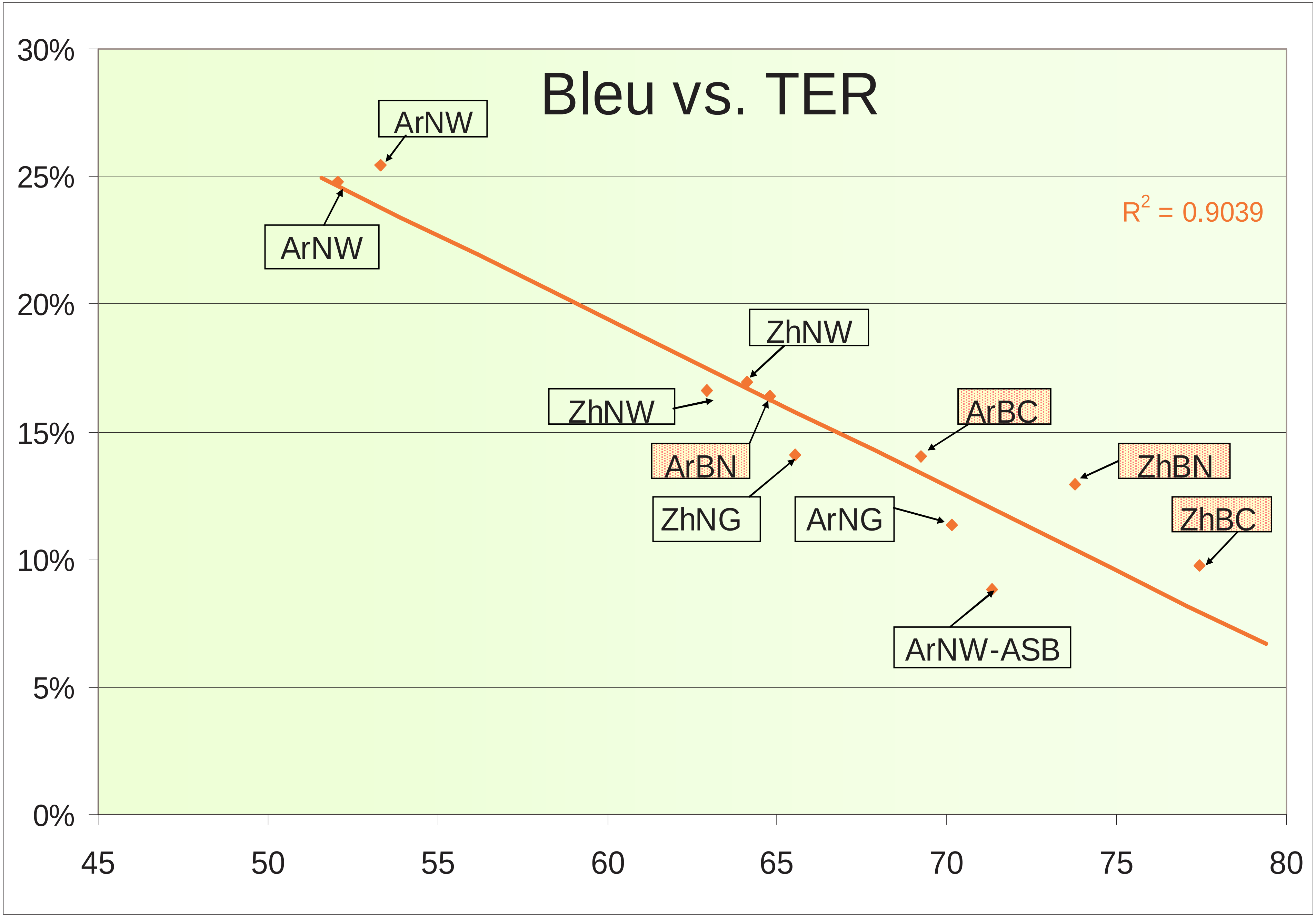
DISTILLATION GNG: Sample NL Question Schemata II

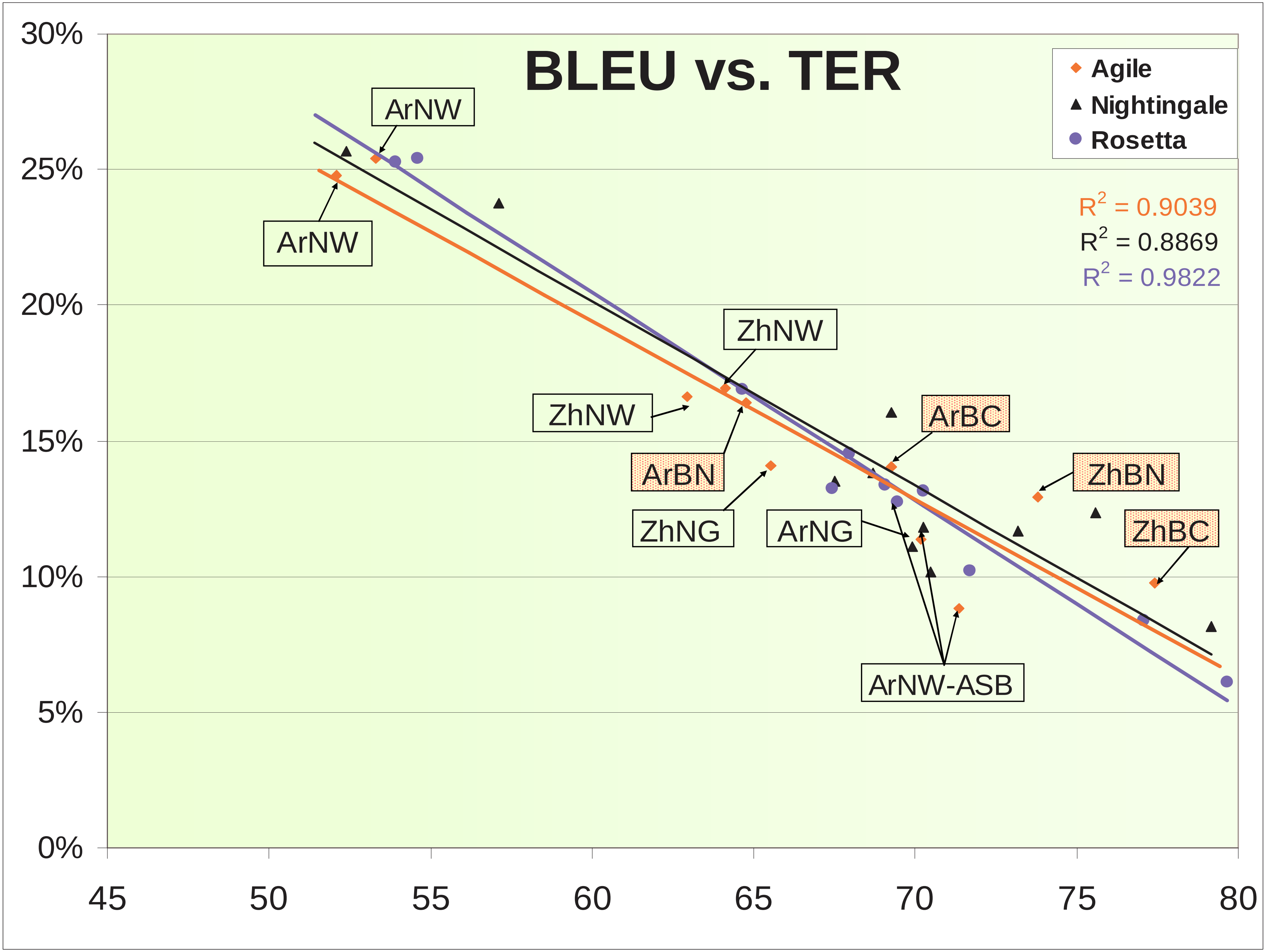
SPECIFIC:

- 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
R2 = 0.9039
30%
25%
20%
15%
10%
5%
0%
45
50
55
60
65
70
75
80
ArNW
ArNW
ZhNW
ZhNW
ArBN
ArBC
ZhBN
ZhNG
ArNG
ZhBC
ArNW-ASB

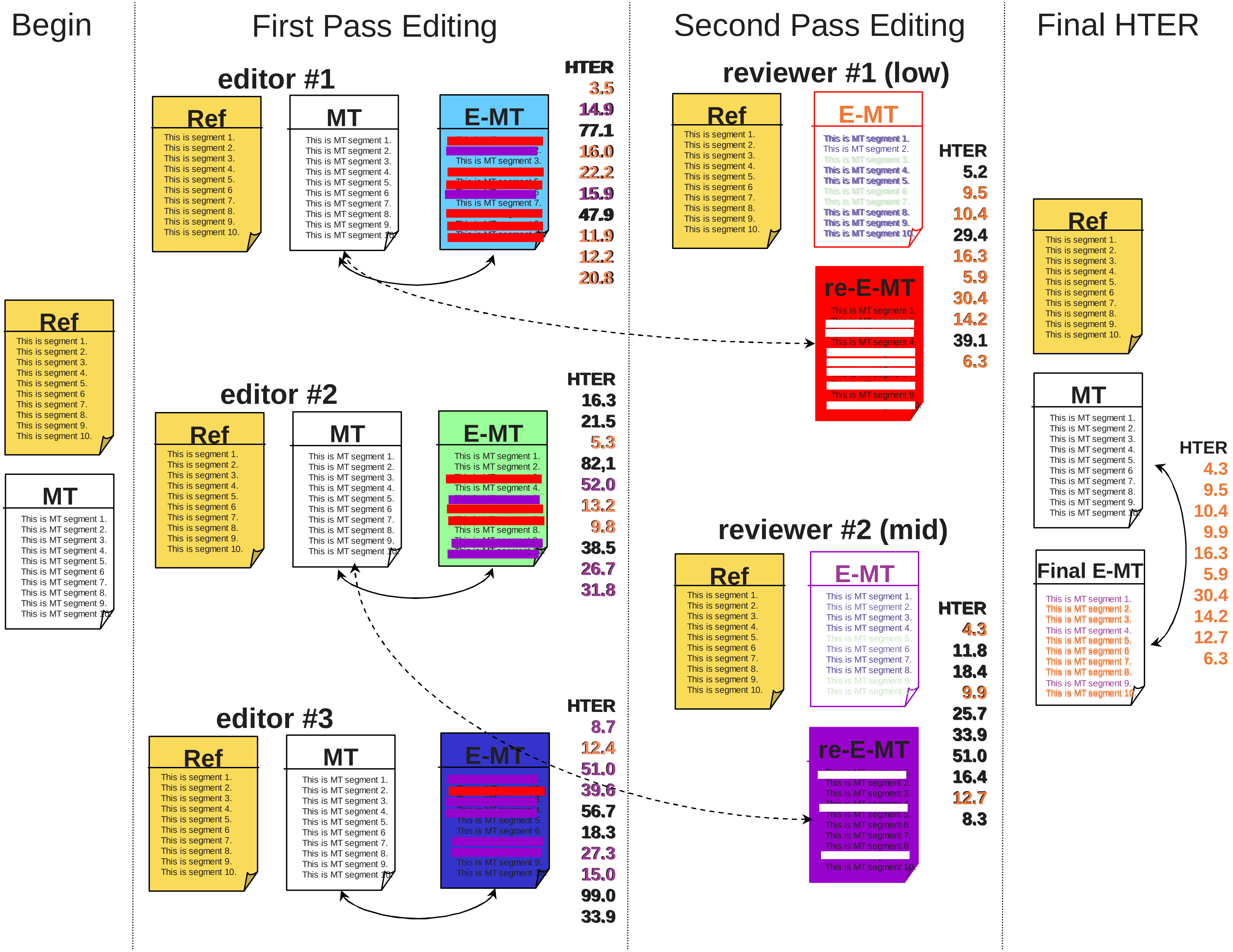

BLEU vs. TER
30%
25%
20%
15%
10%
5%
0%
45
50
55
60
65
70
75
80
Agile
Nightingale
Rosetta
R2 = 0.9039
R2 = 0.8869
R2 = 0.9822
ArNW
ArNW
ZhNW
ZhNW
ArBC
ArBN
ZhBN
ZhNG
ArNG
ZhBC
ArNW-ASB

HTER

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

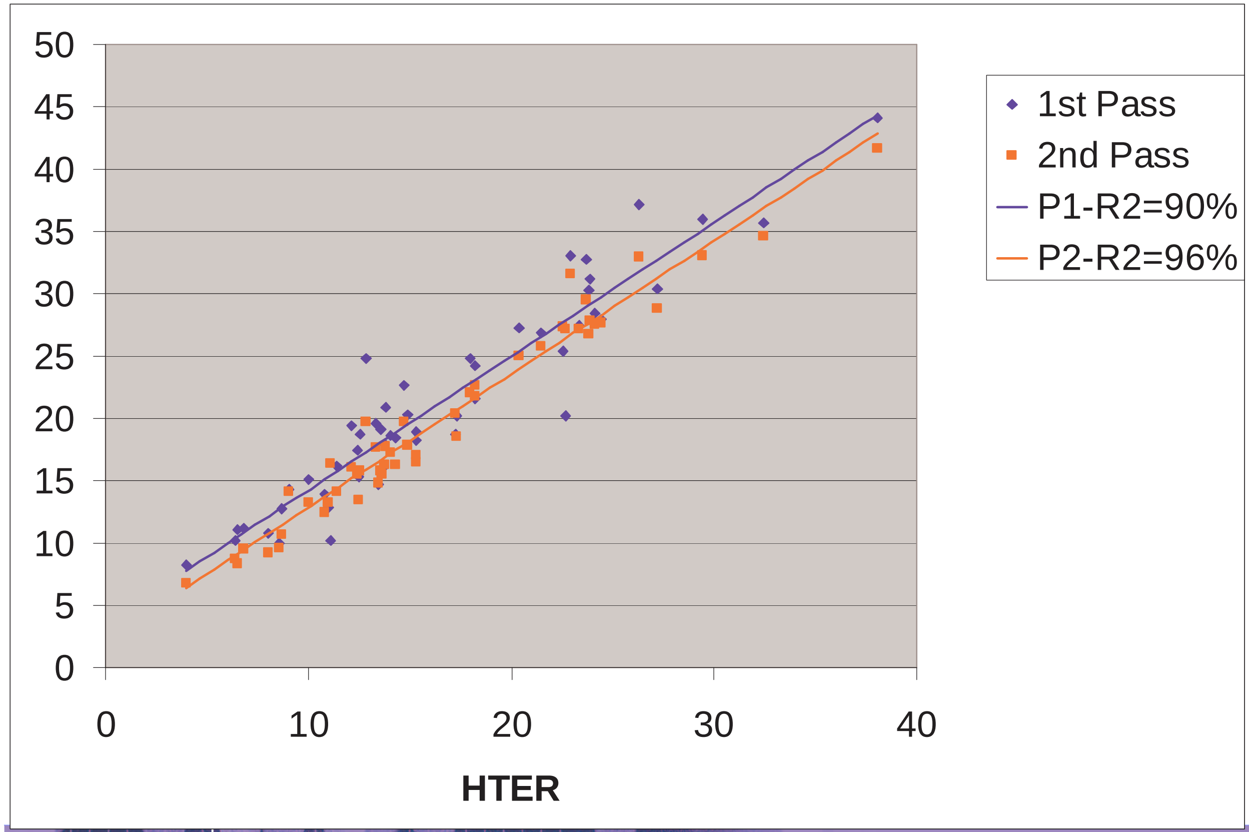
$$HTER = \frac{I + D + S + M}{|R|}$$

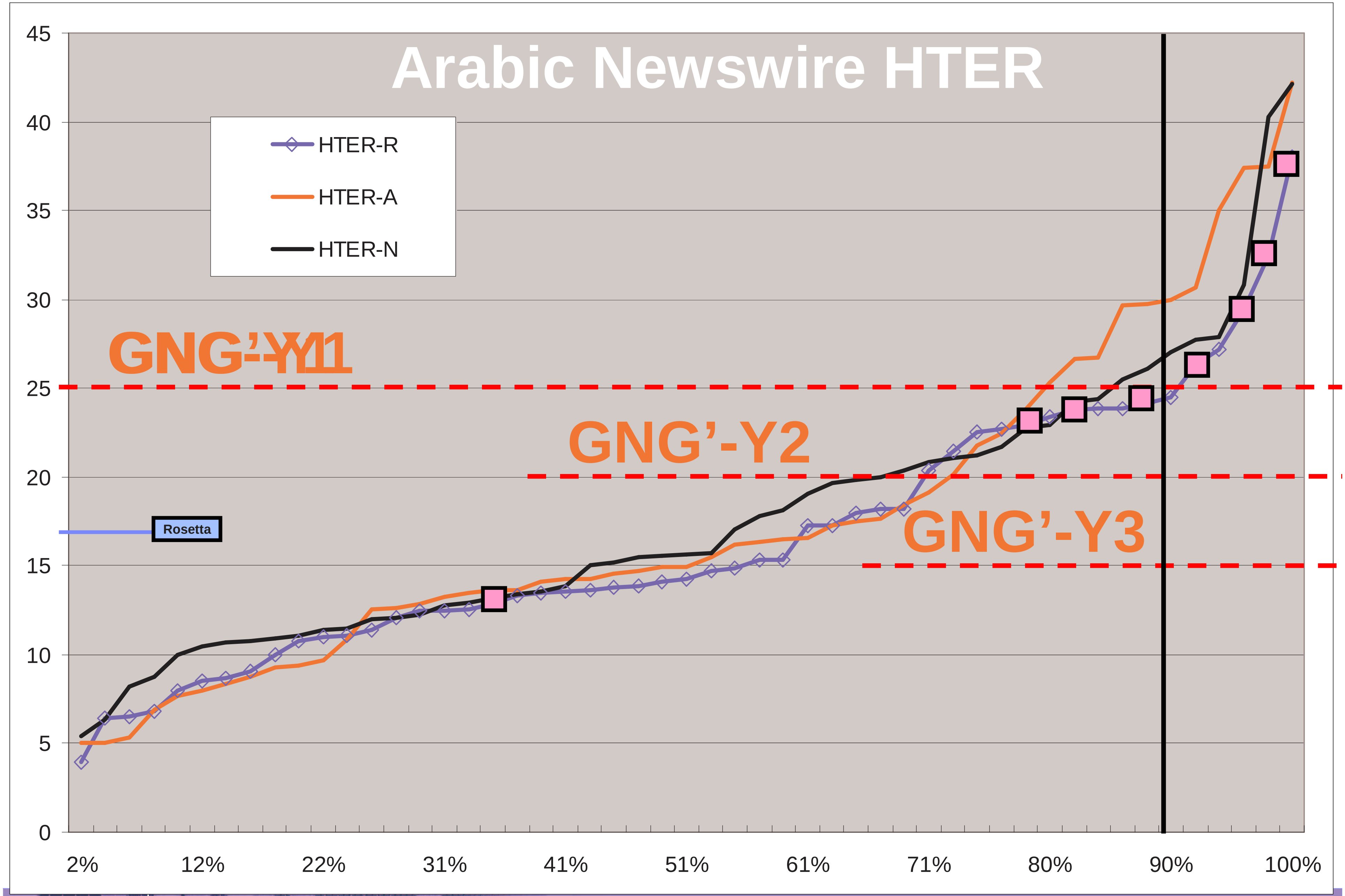
  - M is number of word or phrase shift movements


Begin
First Pass Editing
Second Pass Editing
Final HTER
Ref
This is segment 1.
This is segment 2.
This is segment 3.
This is segment 4.
This is segment 5.
This is segment 6
This is segment 7.
This is segment 8.
This is segment 9.
This is segment 10.
MT
This is MT segment 1.
This is MT segment 2.
This is MT segment 3.
This is MT segment 4.
This is MT segment 5.
This is MT segment 6
This is MT segment 7.
This is MT segment 8.
This is MT segment 9.
This is MT segment 10.
editor #1
Ref
MT
E-MT
This is MT segment 3.
This is MT segment 7.
HTER
3.5
14.9
77.1
16.0
22.2
15.9
47.9
11.9
12.2
20.8
editor #2
Ref
MT
E-MT
This is MT segment 1.
This is MT segment 2.
This is MT segment 4.
This is MT segment 8.
HTER
16.3
21.5
5.3
82,1
52.0
13.2
9.8
38.5
26.7
31.8
editor #3
Ref
MT
E-MT
This is MT segment 5.
This is MT segment 6
This is MT segment 9.
This is MT segment 10.
HTER
8.7
12.4
51.0
39.6
56.7
18.3
27.3
15.0
99.0
33.9
reviewer #1 (low)
Ref
E-MT
re-E-MT
This is MT segment 1.
This is MT segment 4.
This is MT segment 9.
HTER
5.2
9.5
10.4
29.4
16.3
5.9
30.4
14.2
39.1
6.3
reviewer #2 (mid)
Ref
E-MT
re-E-MT
This is MT segment 2.
This is MT segment 3.
This is MT segment 5.
This is MT segment 6
This is MT segment 7.
This is MT segment 8.
This is MT segment 10.
HTER
4.3
11.8
18.4
9.9
25.7
33.9
51.0
16.4
12.7
8.3
Ref
MT
Final E-MT
HTER
4.3
9.5
10.4
9.9
16.3
5.9
30.4
14.2
12.7
6.3

LDC multipass Post Editing

Rosetta	P1	P2	FINAL
NW	21.2%	19.8%	16.5%
Delta		-1.4%	-4.7%
R2	90%	96%	




Arabic Newswire HTER
HTER-R
HTER-A
HTER-N
GNG'-Y1
GNG'-Y2
GNG'-Y3
Rosetta
0
5
10
15
20
25
30
35
40
45
2%
12%
22%
31%
41%
51%
61%
71%
80%
90%
100%

XIN_ARB_20060216.0141 HTER=4%

The French President to Visit India to Intensify Bilateral Cooperation 0

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 $ordinal. 1

It is expected to be the signing of a number of agreements and memoranda of understanding during the visit reflectsing the extent of the cooperation between India and France. 1

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. 0

The two countries aim to intensify bilateral cooperation in various fields, including their partnership in the political, economic, defense, space, and civilian nuclear energy. 1

President Jacques Chirac will deliver a keynote speech on economic partnership between India and France. 0

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. 0

XIN_ARB_20060212.0073 HTER=15.3% BLEU=.25

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

NW TEXT		
STD. ERR. Doc=302wd	TER	BLEU
Agile	5.0	5.7
Nightingale	5.8	5.7
Rosetta	5.3	5.5

BN AUDIO		
STD. ERR. Doc=770wd	TER	BLEU
Agile	4.5	4.9
Nightingale	6.6	4.5
Rosetta	4.2	4.5

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

Post Editing	Agile	Nightingale	Rosetta
LDC HTER	21.01%	20.18%	19.19%
IBM HTER	34.02%	32.94%	32.91%
R2	62%	59%	58%
STD ERR	5.9%	5.0%	5.9%

+65%

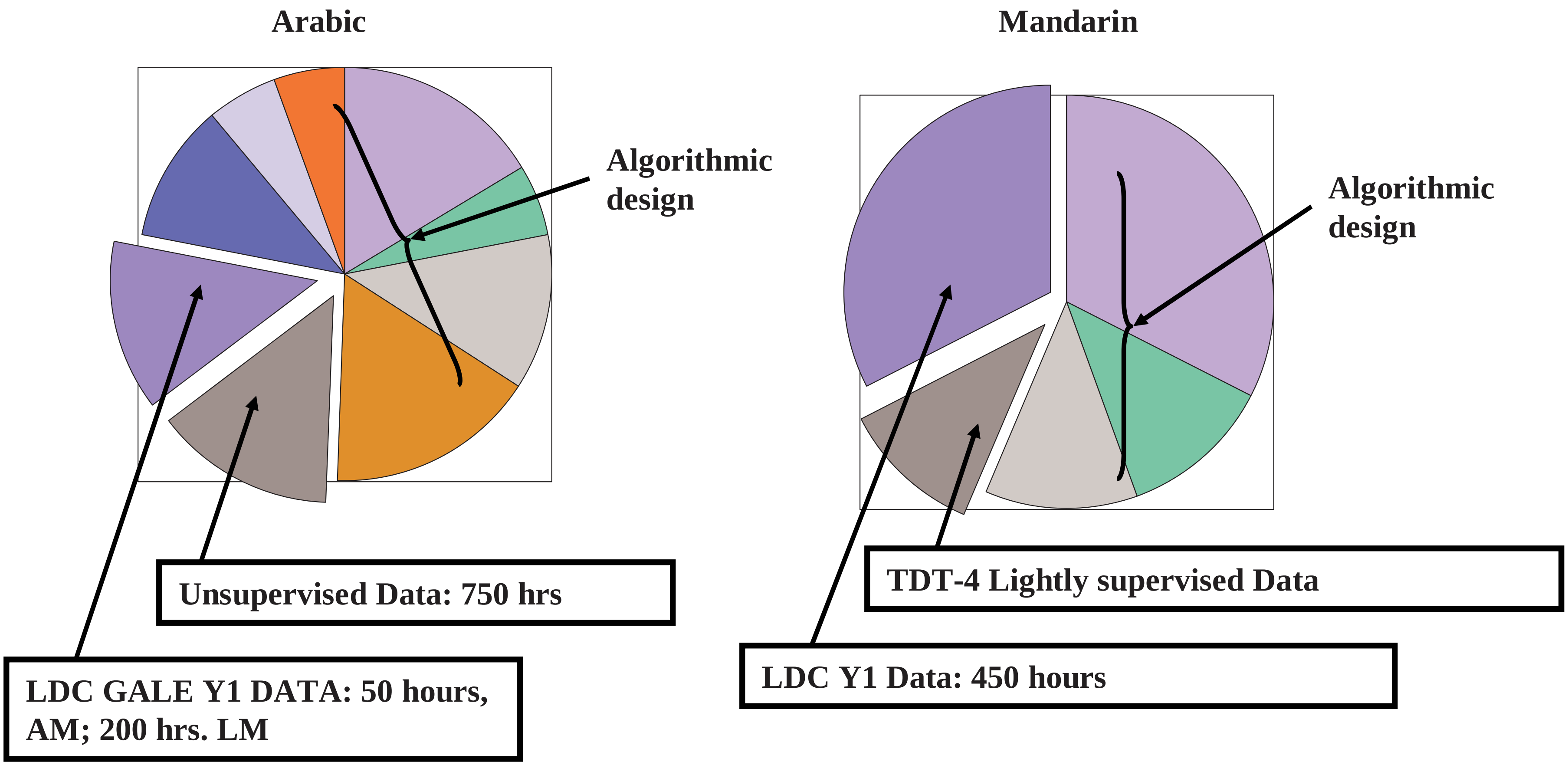
- Similar results for Chinese

The 2006 Rosetta Transcription Effort

Net Rosetta Progress This Year

	Mandarin (RT04 Test set)	Arabic (RT04 Test set)
December	**23.2%**	**21.7%**
June	**13.5%**	**12.6%**
Improvement	**42%**	**42%**

Where did the improvement come from?



Arabic

Mandarin

Algorithmic design

Algorithmic design

Unsupervised Data: 750 hrs

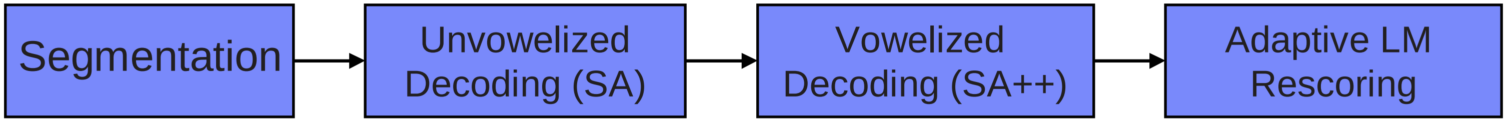
LDC GALE Y1 DATA: 50 hours, AM; 200 hrs. LM

TDT-4 Lightly supervised Data

LDC Y1 Data: 450 hours

Transcription Flow Charts

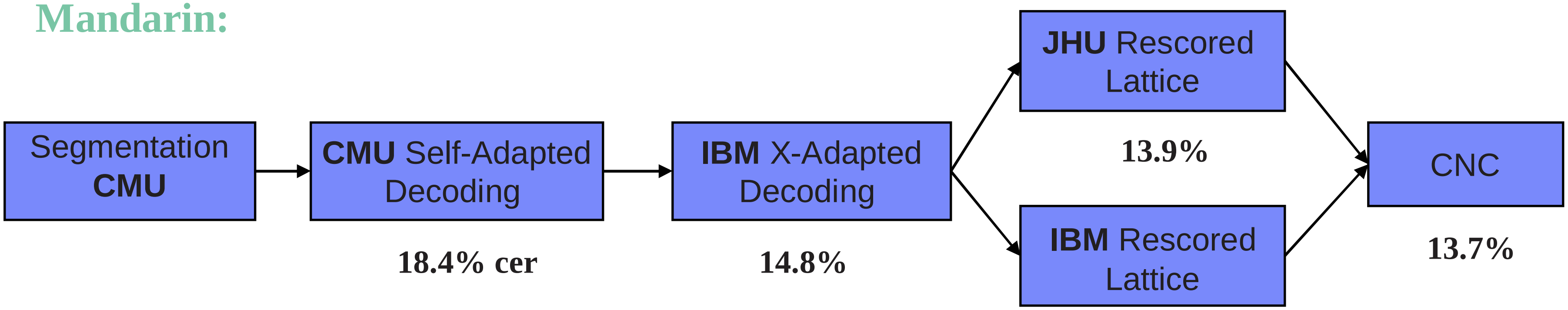
Arabic:



15.3% wer, bnat 13.7% 13.4%

* Numbers on subset of BNAT and BCAD

Mandarin:



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)

Pron. Prob.	RT-04	BNAT-05	BCAD-05
no	16.0%	17.3%	26.0%
yes	14.9%	16.4%	25.1%

- Developed technology to cope with 2 million pronunciations
- Significant improvements from pronunciation probabilities

Vowelization and Broadcast Conversations ..

- ML models : VTLN, FMLLR, MLLR

	RT-04	BNAT-05	BCAD-05
Unvowelized	17.0%	18.7%	25.4%
Vowelized	14.9%	16.4%	25.1%

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

Evaluation Results

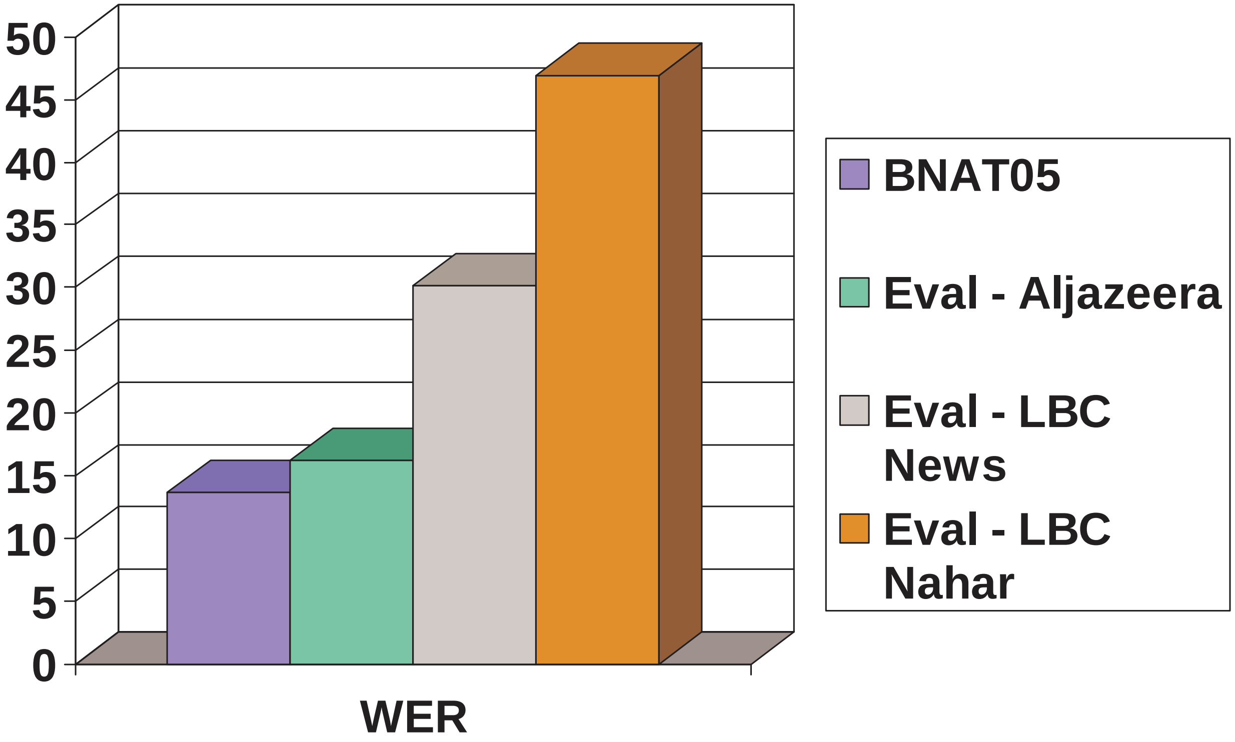
	BC	BN
Arabic - Dev	21.5	13.7
- Test	34.0	24.4
- HTER	35.6	29.2
Mandarin - Dev	20.7	12.9
- Test	24.1	13.4
- HTER	37.1	32.4

Really big mismatch between dev & test

We hit the target!

Some mismatch between dev & test

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

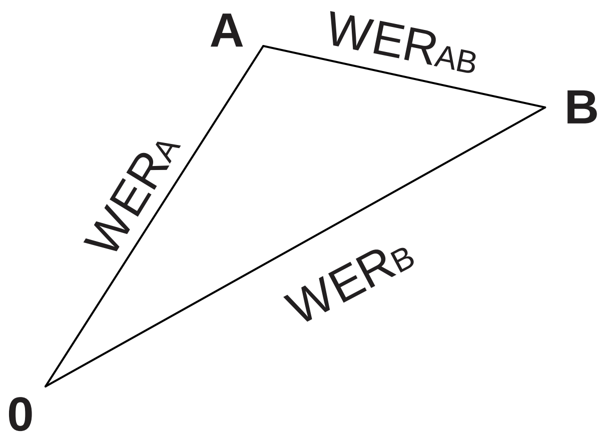
- Rapidly assess the performance of an ASR system on a new test set without the need of a reference transcript

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

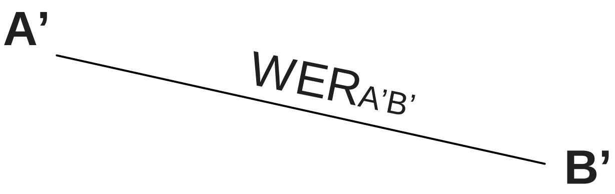
- Applications
  - Predict system performance in government evaluations ☺
  - Select data for (un)supervised training (active learning)
  - Change system configuration to minimize predicted WER

How can we compute $\mathrm{WER}_{A'}$?

Training: all WERs known

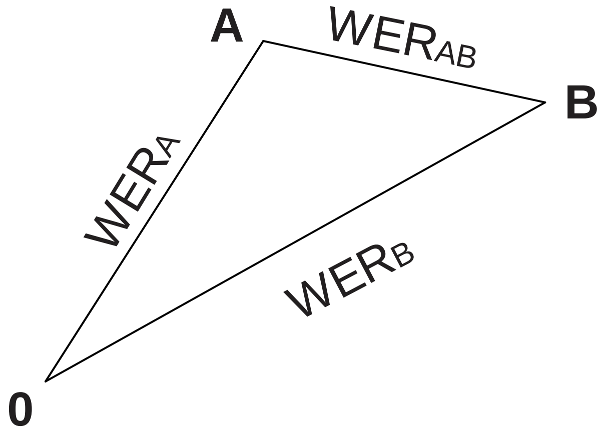


Test: only $\mathrm{WER}_{A'B'}$ known

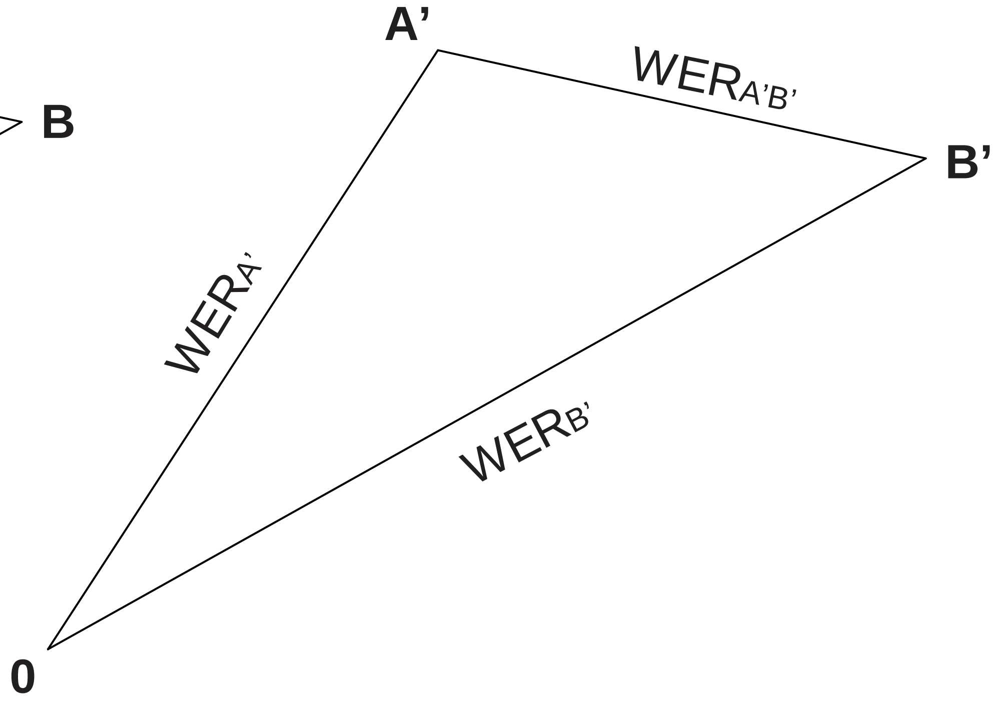


How can we compute $\mathrm{WER}_{A'}$?

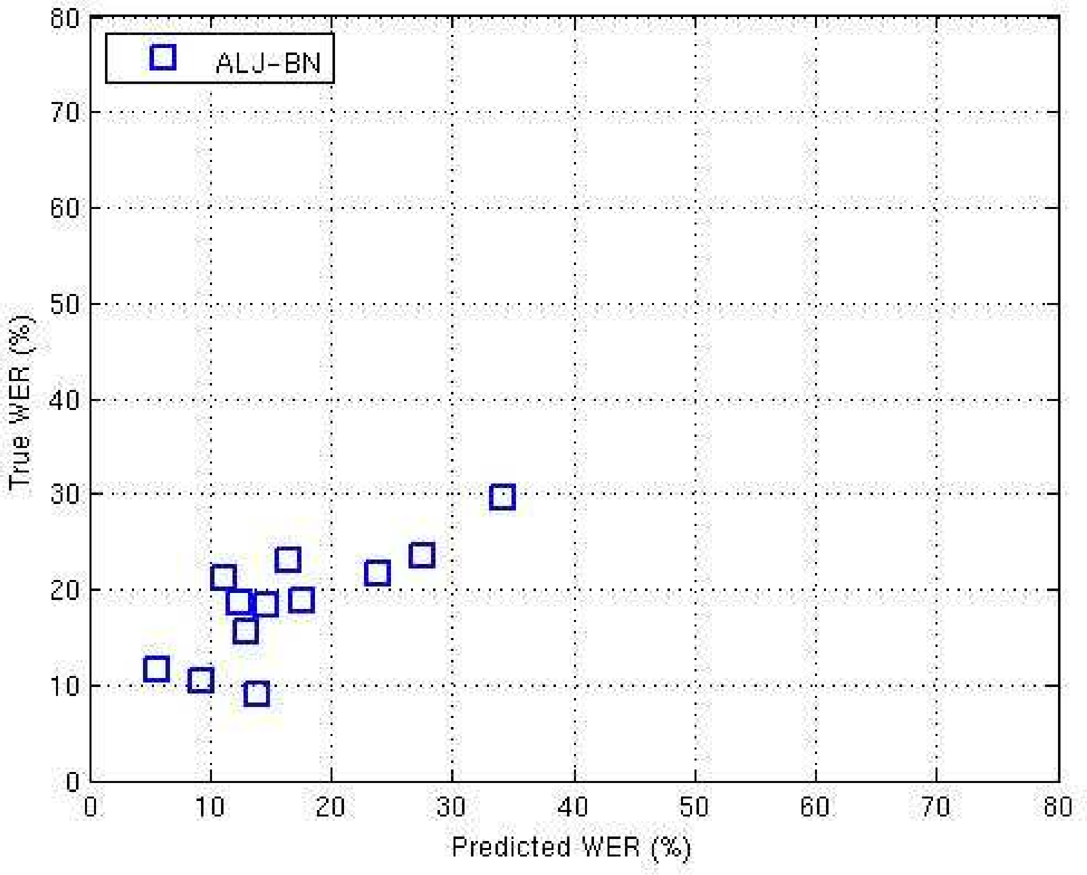
Training: all WERs known



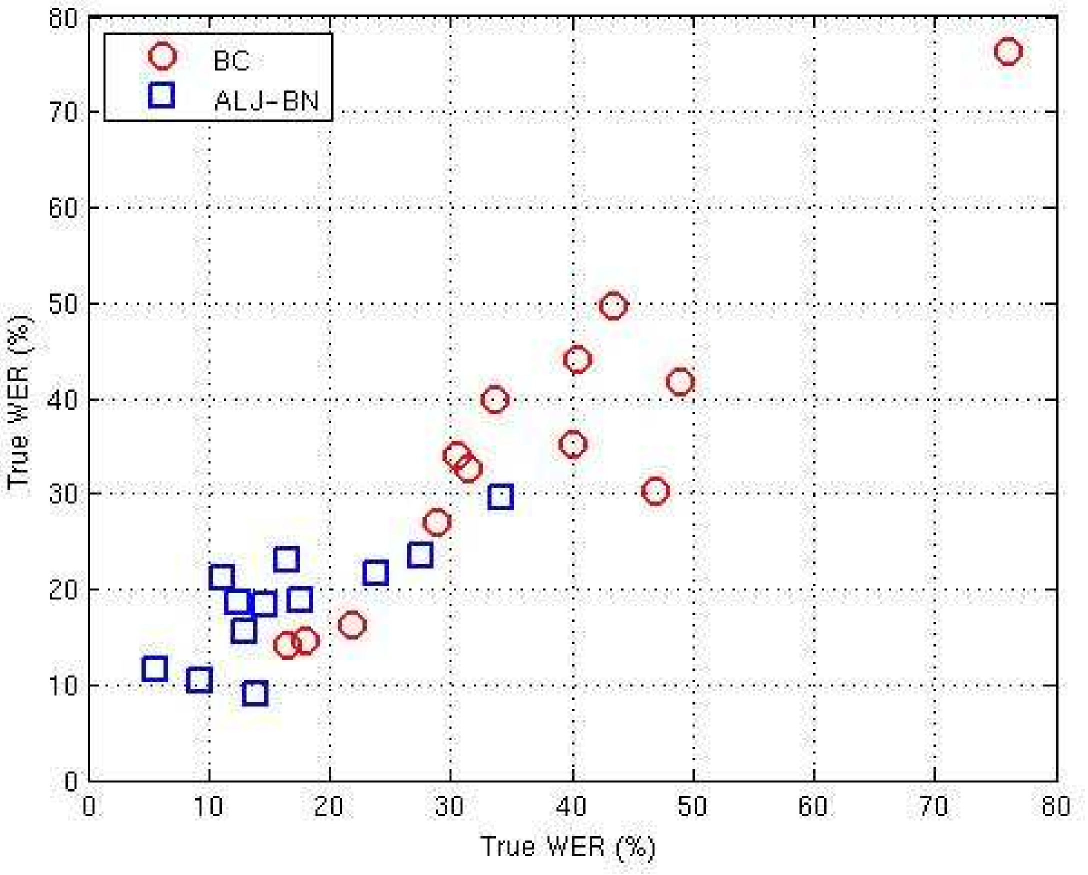
Test: only $\mathrm{WER}_{A'B'}$ known



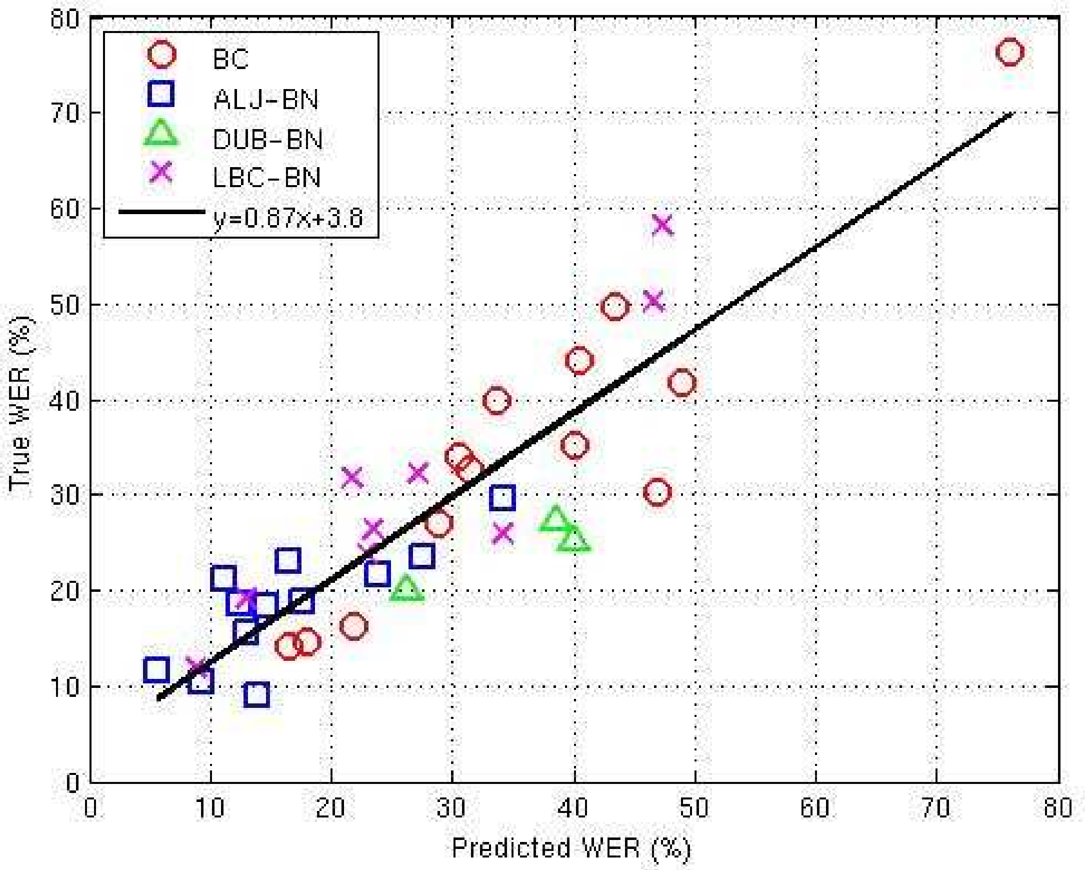
Performance on the 2006 GALE evaluation data



Performance on the 2006 GALE evaluation data



Performance on the 2006 GALE evaluation data

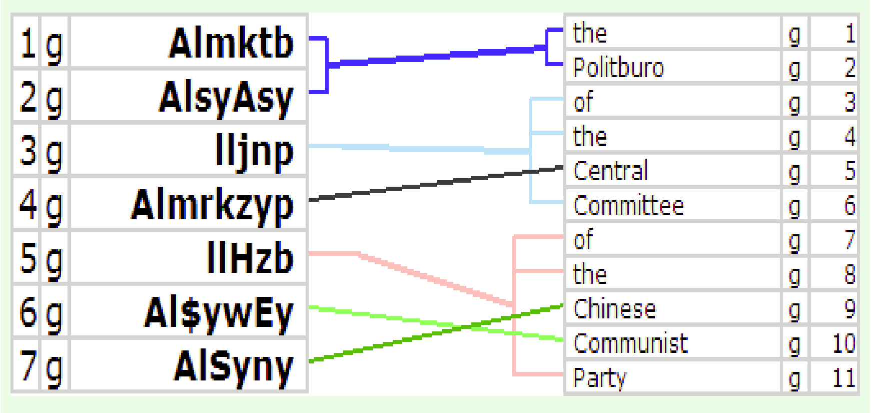


True WER=29.2%, predicted WER=30.0%, CORR=0.87, MAD=5.4

Rosetta:
MT GALE GnG06 Report

A Direct Translation Model II

How many phrases do we need?

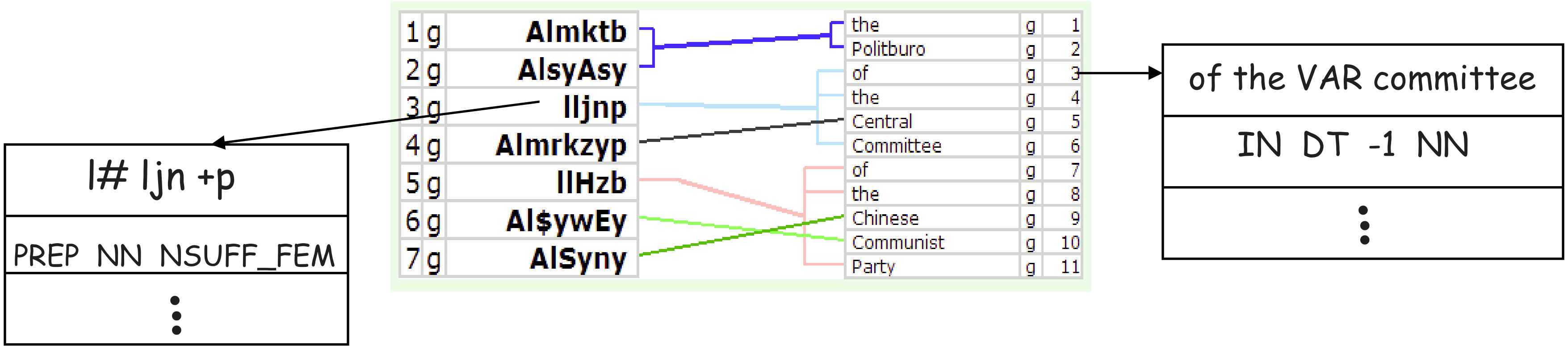


- N-M blocks (Used by most SMT systems)
  - General
    - All possible blocks extracted
    - 40-50M blocks in Arabic
    - Sparsity problems

lljnp	Almrkzyp
committee of the commission commission of the committee the committee of the commission on the commission committee of	central the central ***of the central*** of central and the central and central , central 's central

of the central committee (11)
of the central committee of (11)
the central committee of (8)
central committee (7)
committee central (2)

DTM Decoder (aka MaxEnt)



1	g	Almktb
2	g	AlsyAsy
3	g	lljnp
4	g	Almrkzyp
5	g	llHzb
6	g	Al$ywEy
7	g	AlSyny

the	g	1
Politburo	g	2
of	g	3
the	g	4
Central	g	5
Committee	g	6
of	g	7
the	g	8
Chinese	g	9
Communist	g	10
Party	g	11

l# ljn +p

PREP NN NSUFF_FEM

⋮

of the VAR committee

IN DT -1 NN

⋮

- **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***

lljnp → of the VAR committee

Almrkzyp → central

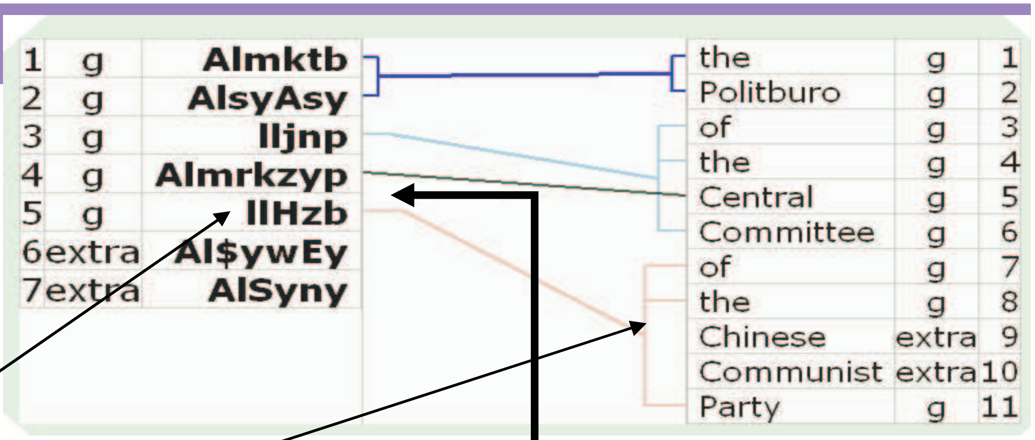
Direct Translation Model

- **Joint future: Jump, Target Sequence**

$$p(T, j \mid S)$$

  - 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



1	g	Almktb
2	g	AlsyAsy
3	g	lljnp
4	g	Almrkzyp
5	g	llHzb
6	extra	Al$ywEy
7	extra	AlSyny

the	g	1
Politburo	g	2
of	g	3
the	g	4
Central	g	5
Committee	g	6
of	g	7
the	g	8
Chinese	extra	9
Communist	extra	10
Party	g	11

- **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**

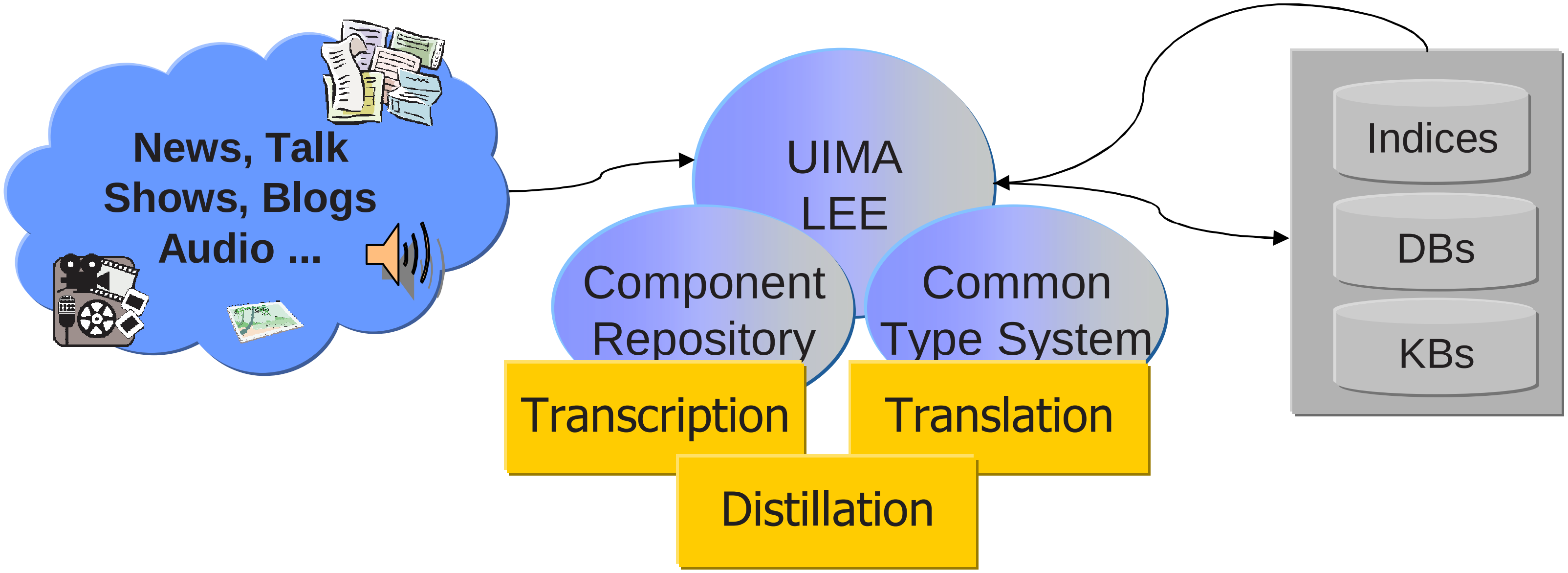
Cnt	Alpha	Jump	Tgt	Seg
1107	1.047	-2	of	l#
3120	0.989	-1	of	l#
55461	**1.319**	**1**	**of**	**l#**
7009	1.225	2	of	l#

- **Block Context Feature**
  - 11 1.66021 0.0330579 1024 -1 party llHzb || communist Al$ywEy chinese AlSyny
- **New Feature ~ coding time + 8 hours training + 1 hr decode time**

Experiments - NIST

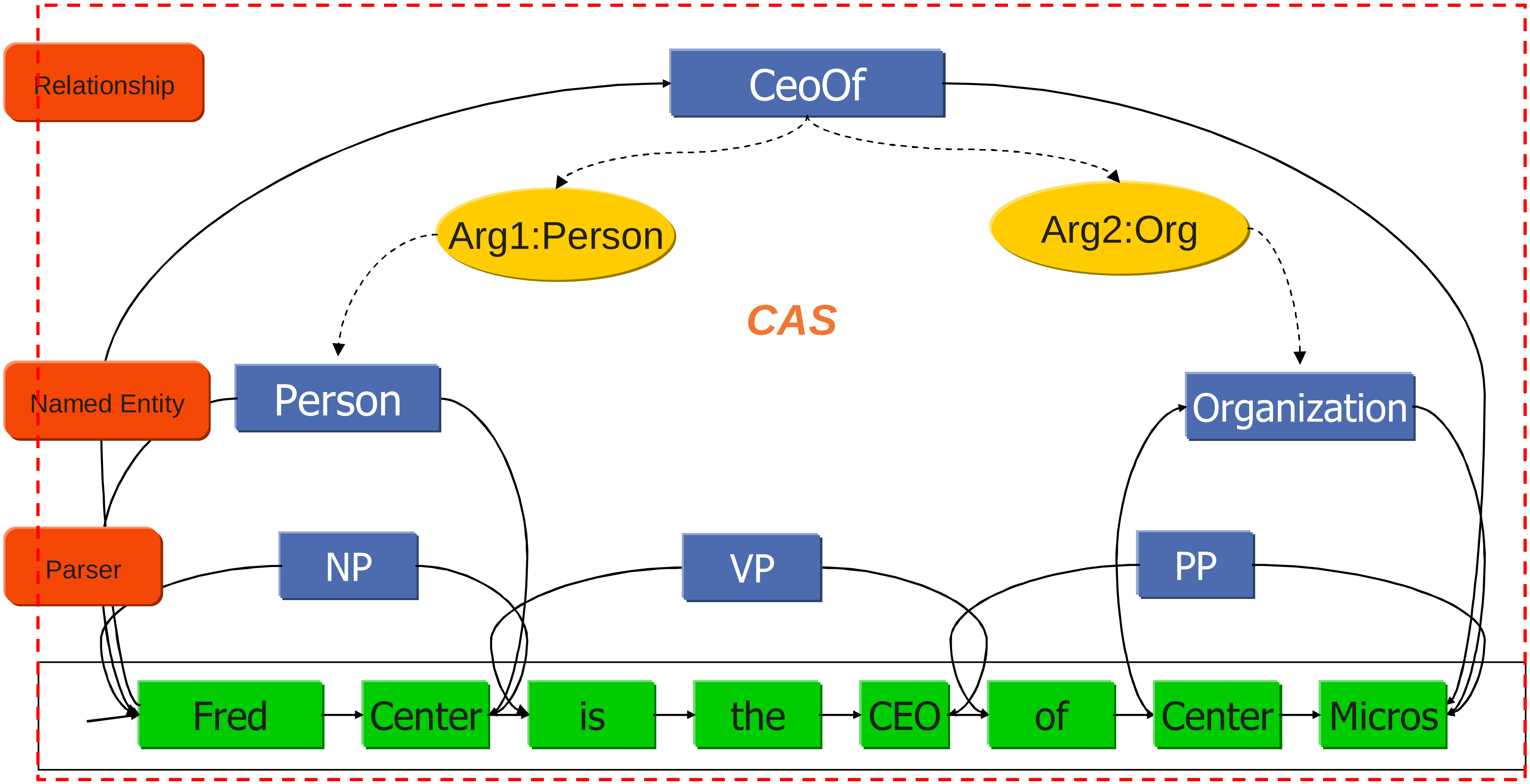
Feature Types	# of feats (MT05)	MT-05	MT-06 (NIST)
MaxEnt Decoder Lexical Feats	520,210	48.21	
+Lexical Context	1,551,582	49.24	
+Segmentation Feats	3,063,023	49.51	
+Part-of-Speech Feats	3,370,901	49.87	
+Distortion Feats	3,412,210	*49.98*	*38.61*
Block Decoder		49.06	36.92

UIMA: ARCHITECTURE FOR DARPA GALE

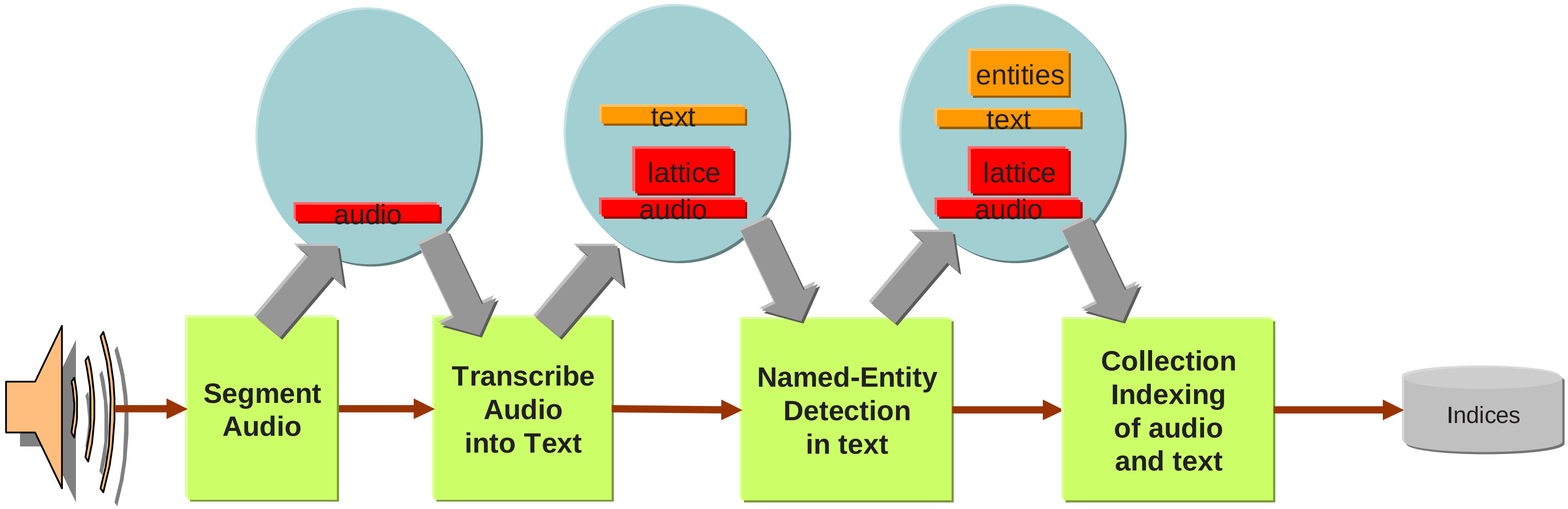


- Highly-distributed plug-and-play architecture
- Support for multi-modal sources
- Support for local/remote heterogenous components
- Open Source

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.

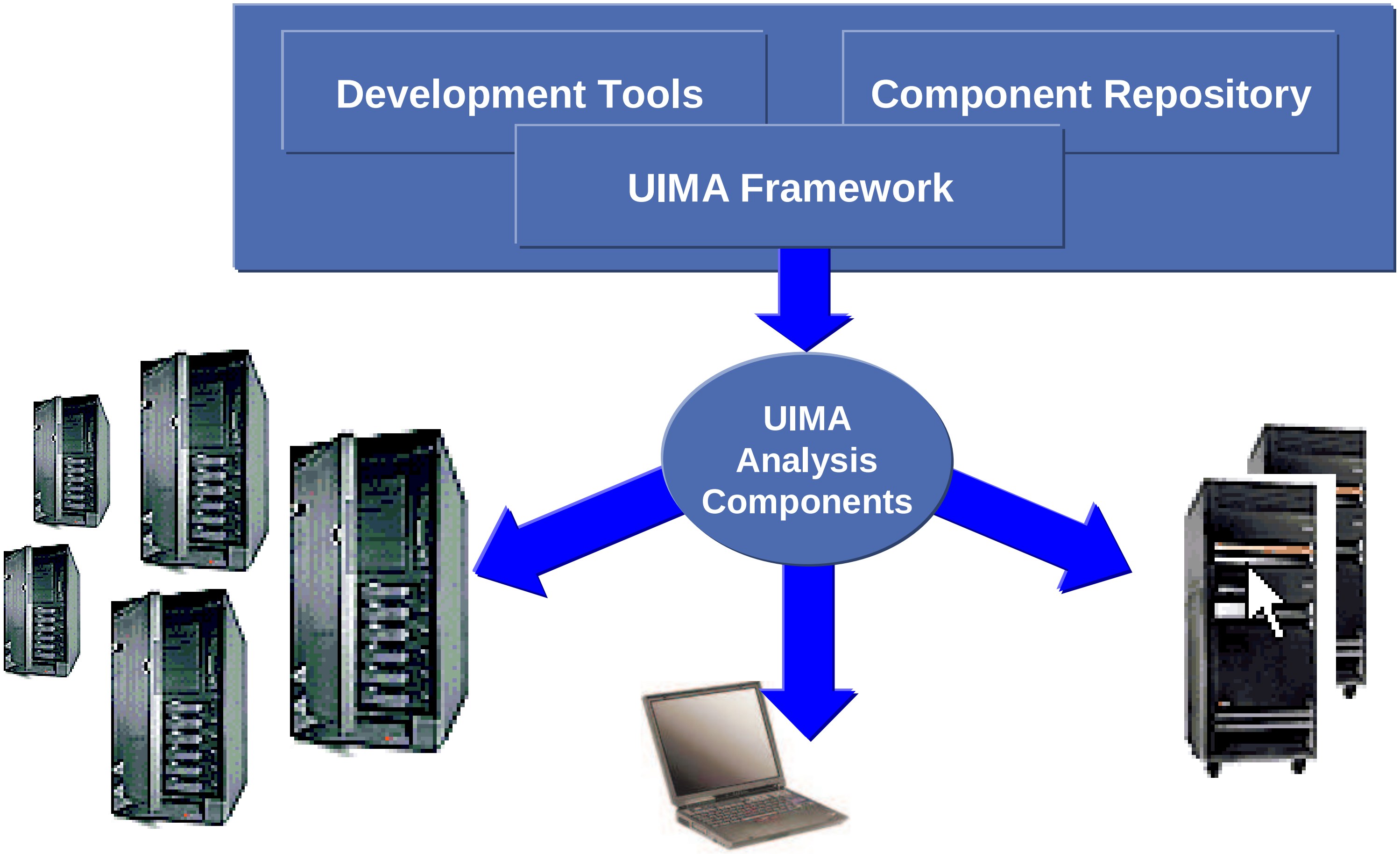


Common Annotation Structure (CAS):
Multiple Subject of Analysis (SOFA) in CAS Supports Multi-Modal Analysis


audio
text
lattice
audio
entities
text
lattice
audio
Segment Audio
Transcribe Audio into Text
Named-Entity Detection in text
Collection Indexing of audio and text
Indices

- 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.



Large-Scale Distributed Deployment

Standalone

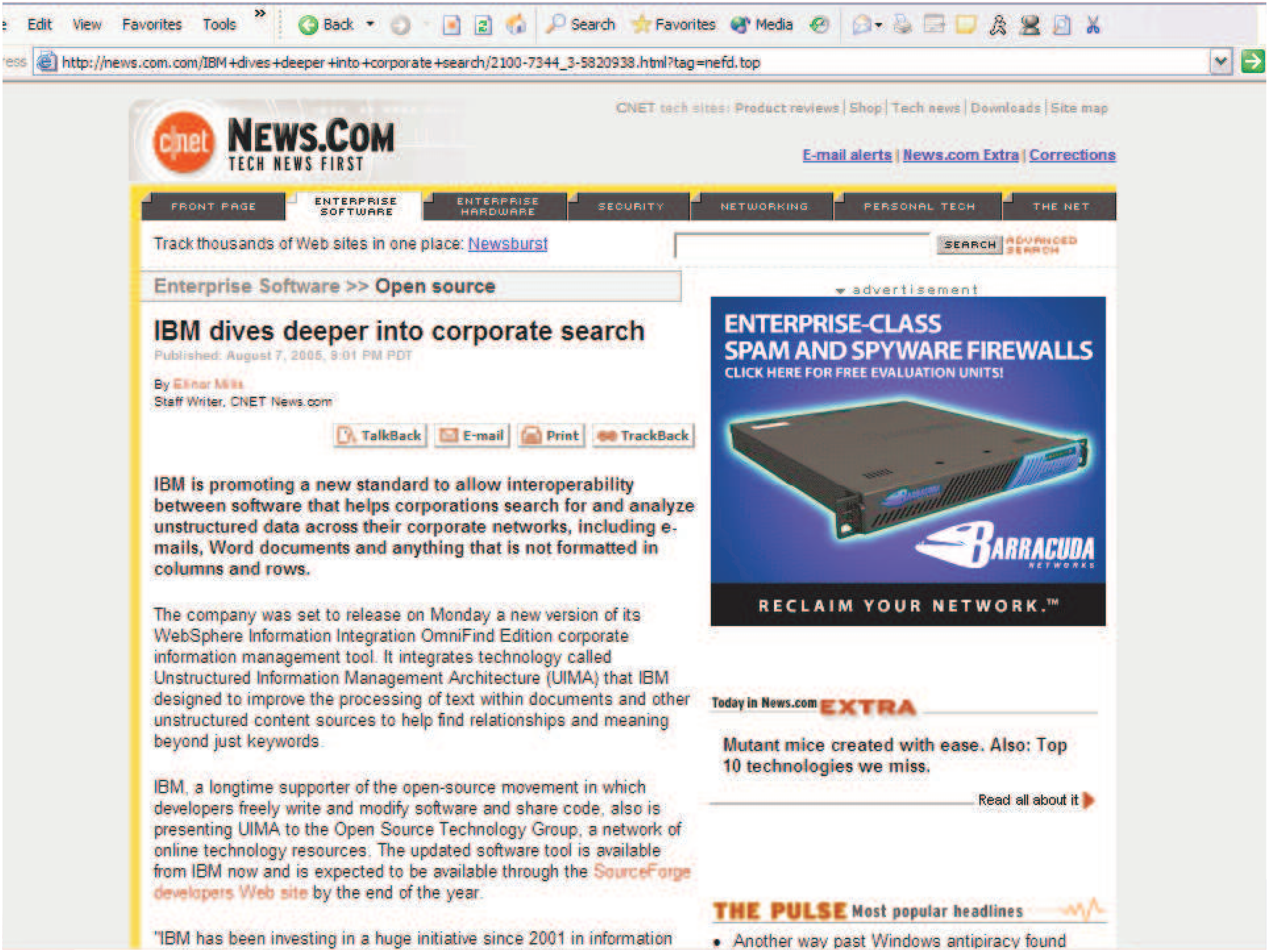
Product Embeddings

Web Fountain, ….

UIMA SDK

OmniFind, WPS, LWP…

Powered by UIMA-Compliant Run-Times



http://news.com.com/IBM+dives+deeper+into+corporate+search/2100-7344_3-5820938.html?tag=nefd.top

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

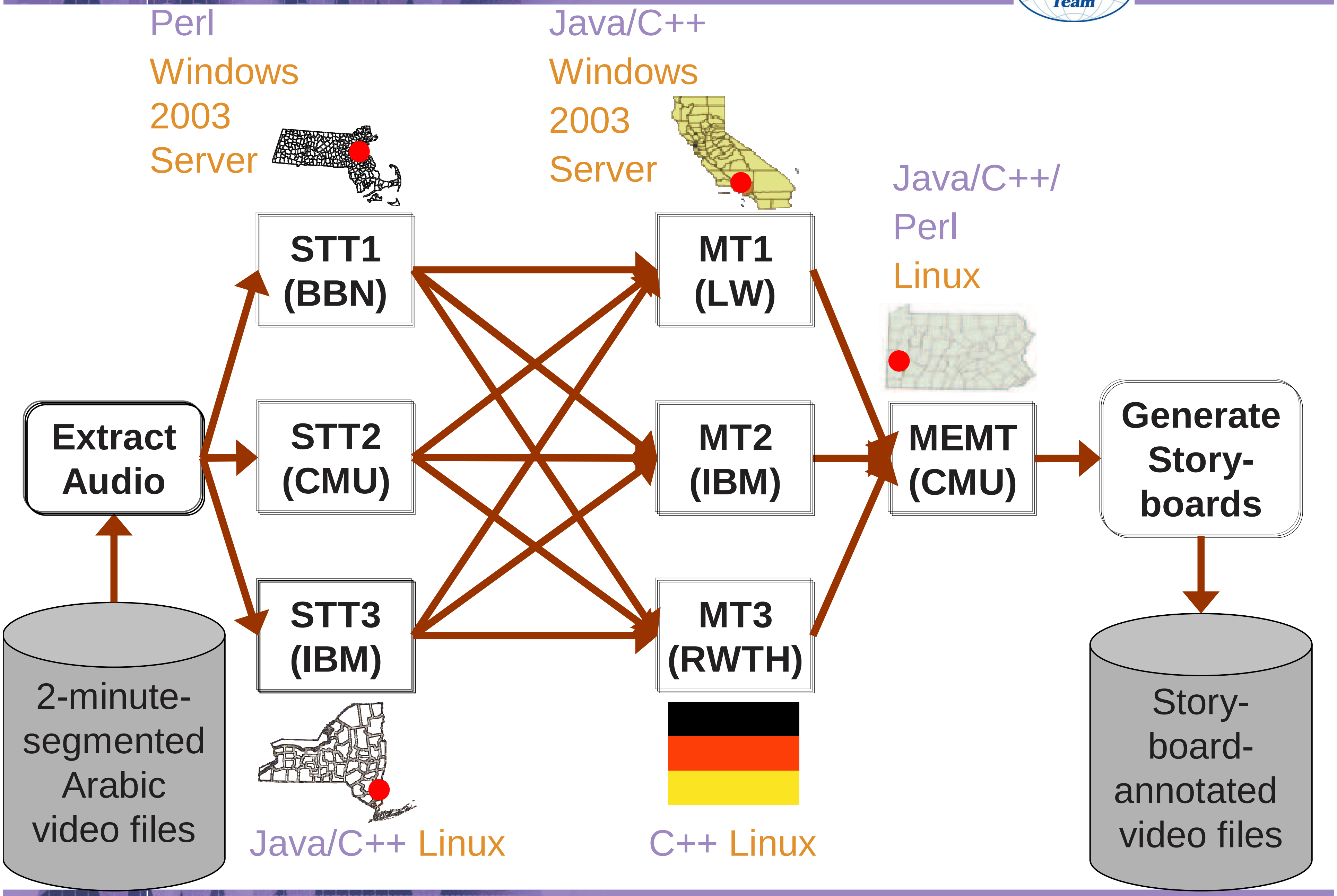
Today in News.com EXTRA

Mutant mice created with ease. Also: Top 10 technologies we miss.

Read all about it

THE PULSE Most popular headlines

- Another way past Windows antipiracy found



Perl Windows 2003 Server

Java/C++ Windows 2003 Server

Java/C++/ Perl Linux

Extract Audio

STT1 (BBN)

STT2 (CMU)

STT3 (IBM)

MT1 (LW)

MT2 (IBM)

MT3 (RWTH)

MEMT (CMU)

Generate Story-boards

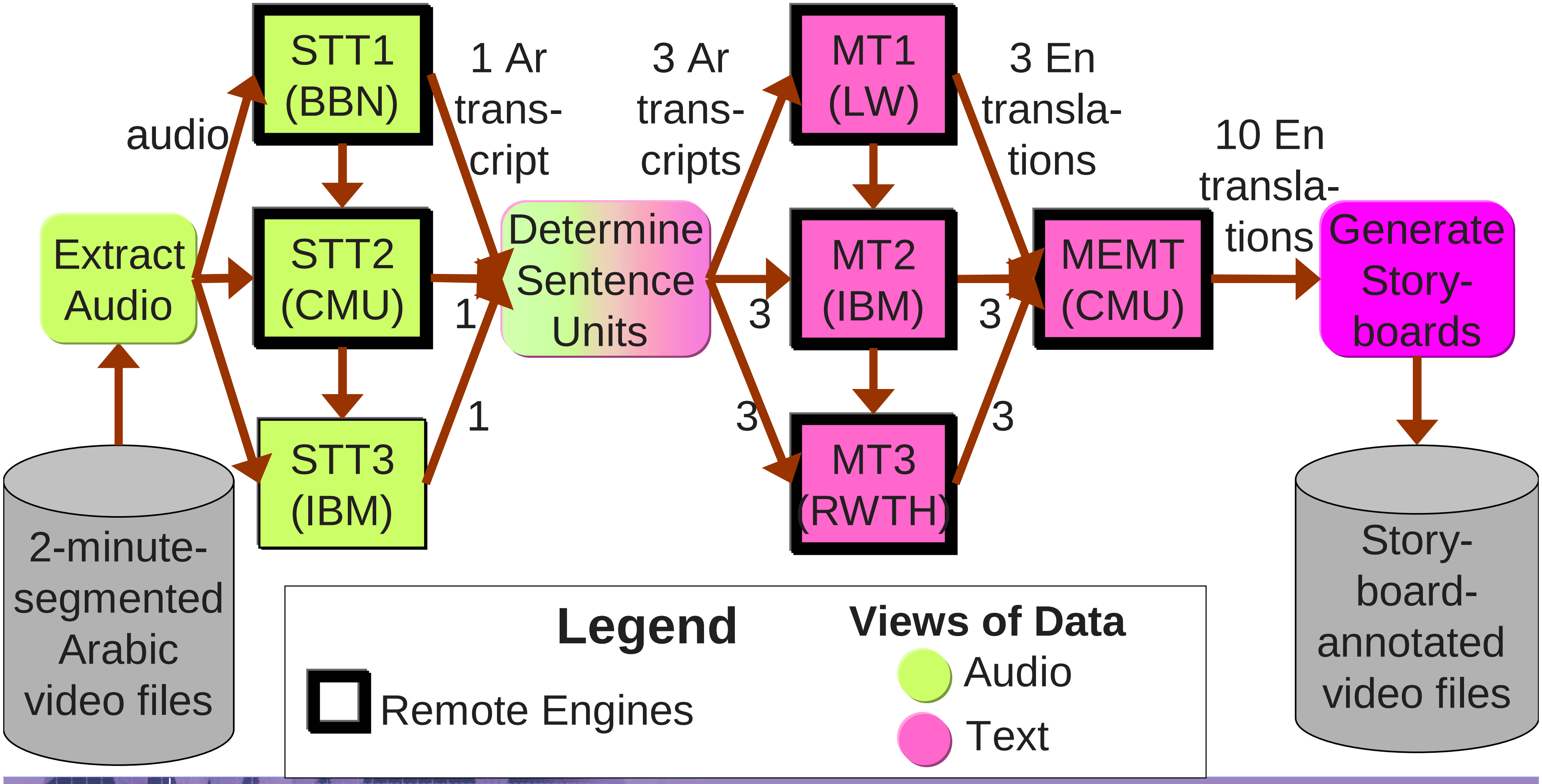
2-minute-segmented Arabic video files

Story-board-annotated video files

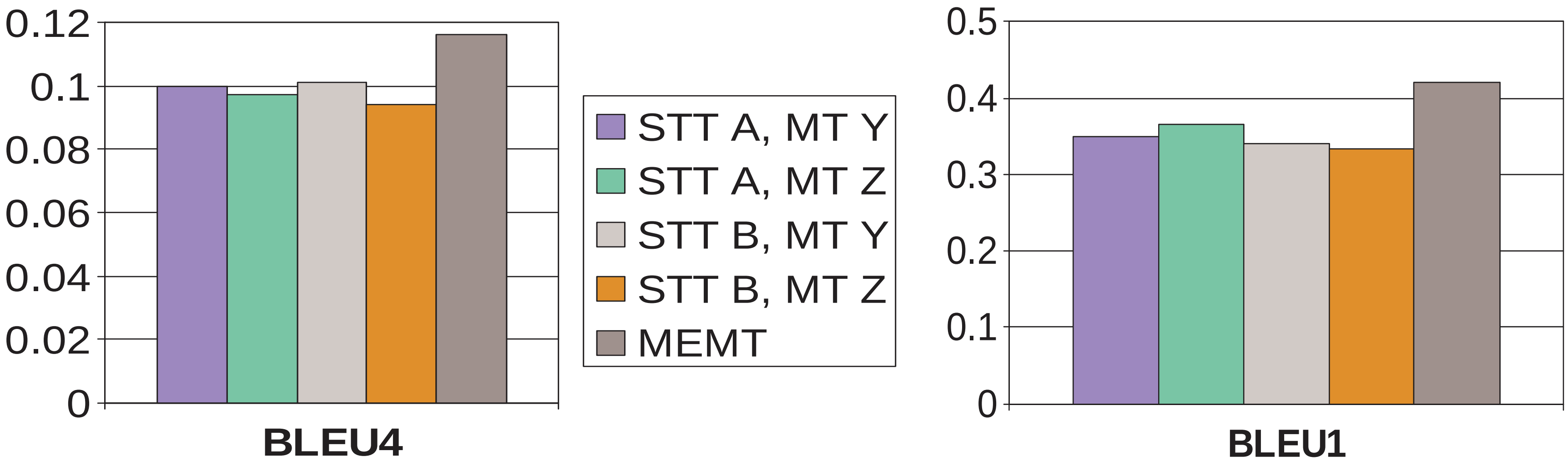
Java/C++ Linux

C++ Linux

Process Data Flow: Serial



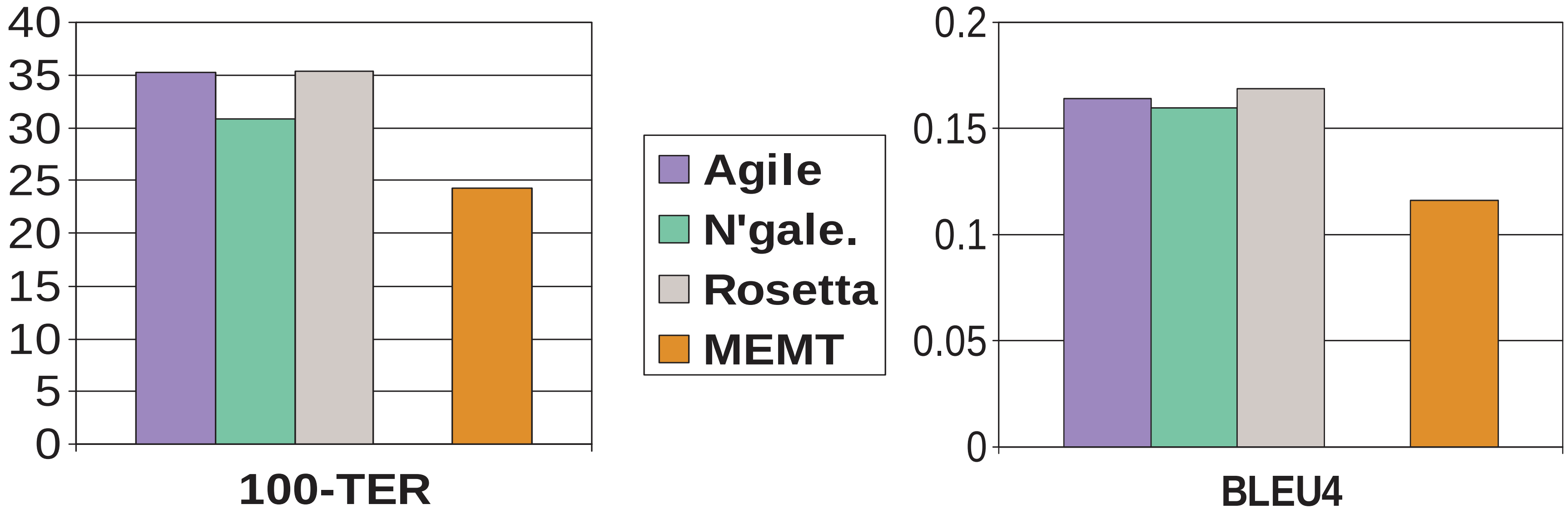
IOD Enables On-Line MEMT, Increased Accuracy



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

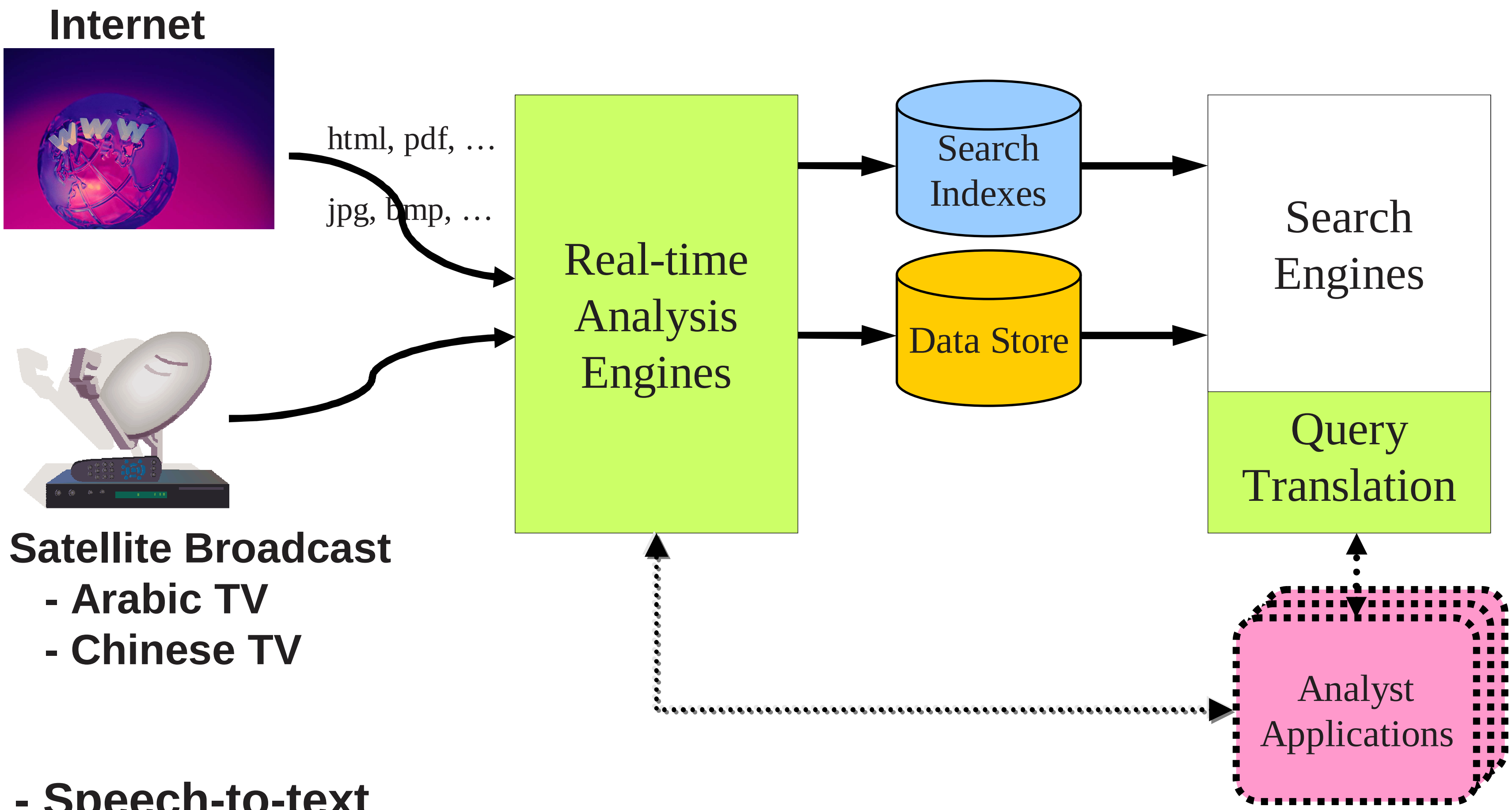
System	TER	BLEU4	BLEU1	METEOR
STT A, MT Y	75.9	0.100	0.349	0.405
STT A, MT Z	75.4	0.097	0.366	0.396
STT B, MT Y	74.7	0.101	0.340	0.405
STT B, MT Z	74.7	0.094	0.334	0.395
MEMT	75.7	0.116	0.421	0.440
MEMT % gain	-1	+15	+15	+9

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

Satellite Broadcast
- Arabic TV
- Chinese TV

- Speech-to-text

- Statistical machine translation

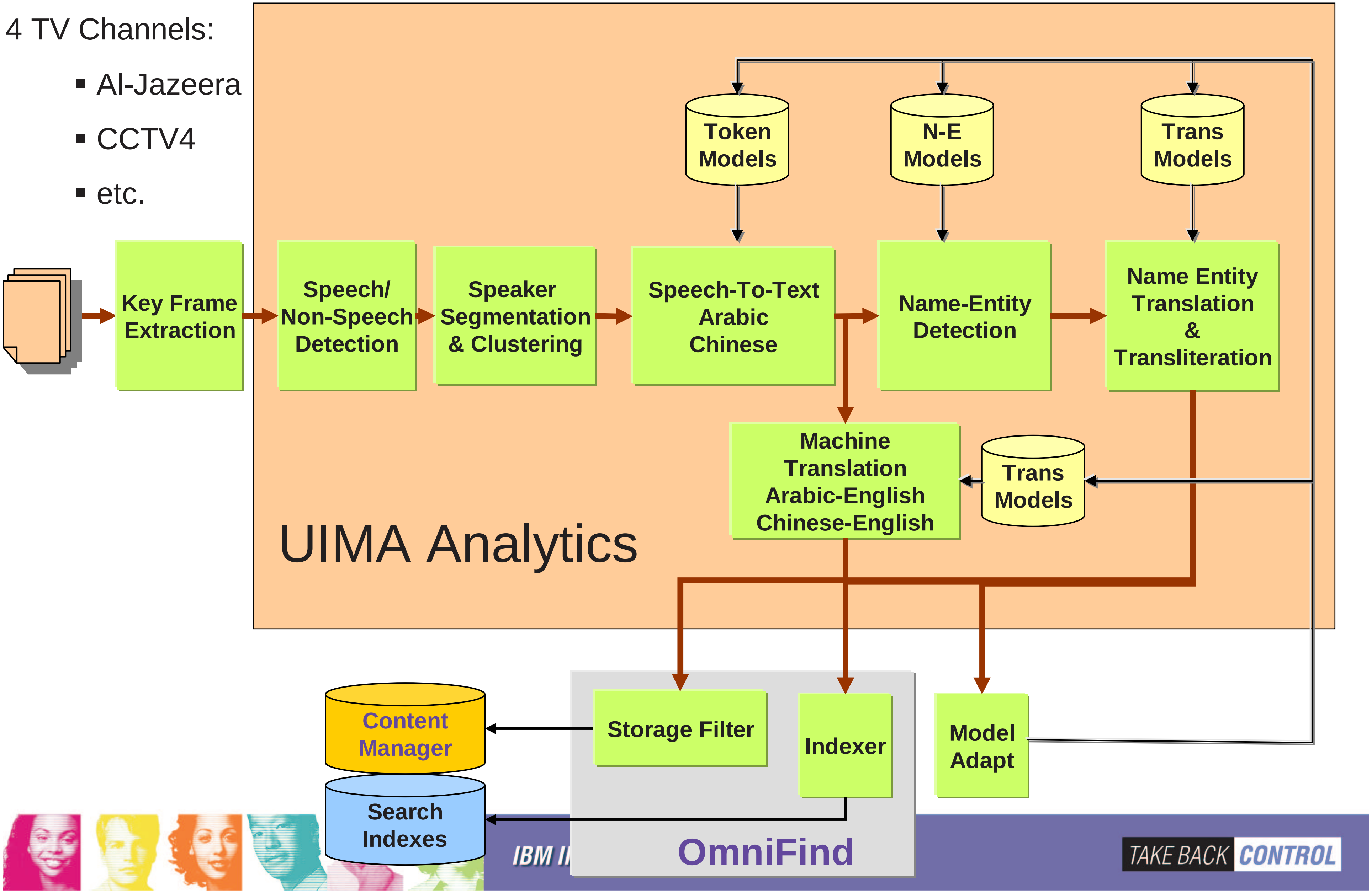
- Cross-lingual search

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

Video Processing Flow

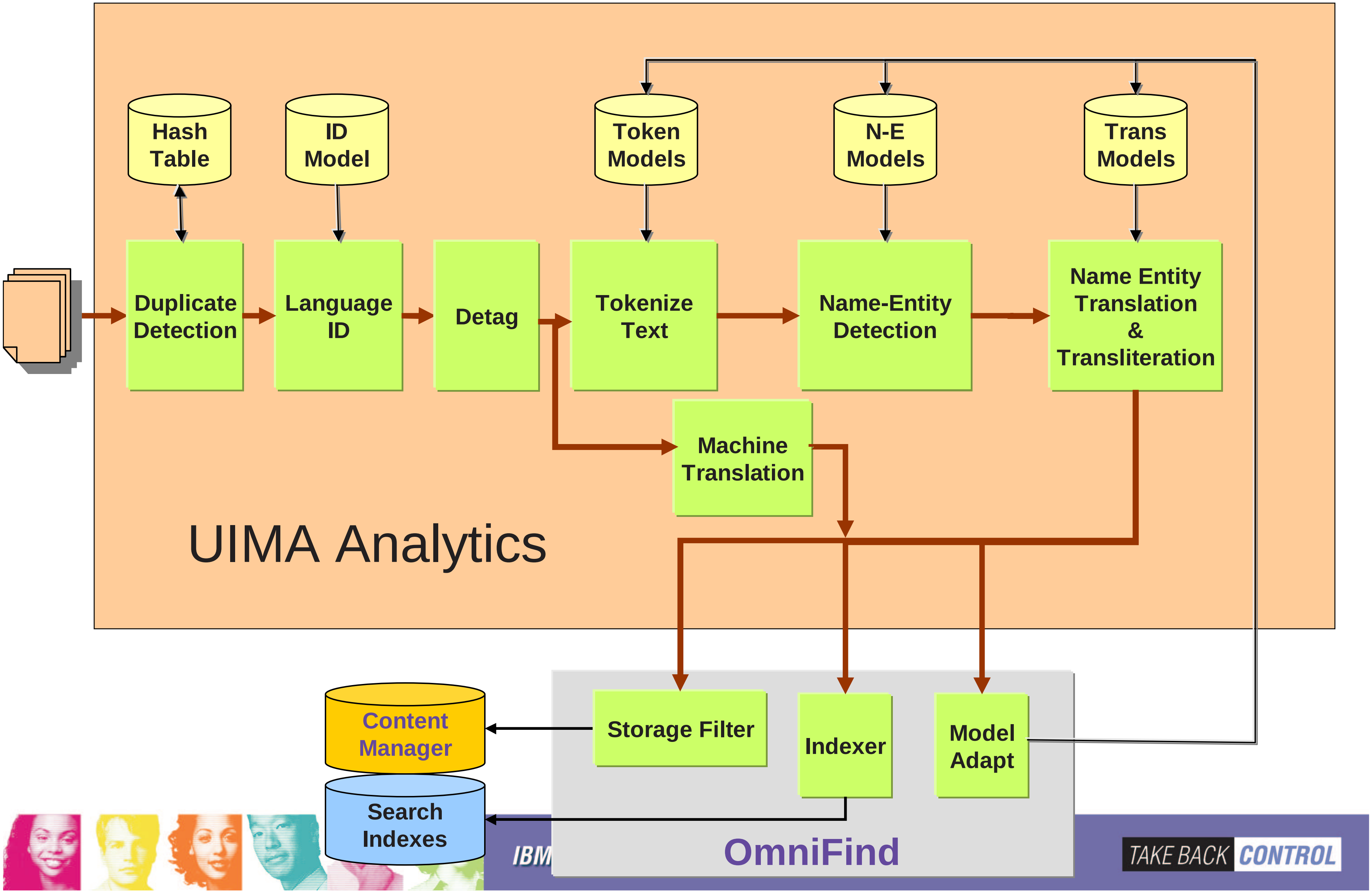
4 TV Channels:

- Al-Jazeera
- CCTV4
- etc.



Token Models
N-E Models
Trans Models
Key Frame Extraction
Speech/ Non-Speech Detection
Speaker Segmentation & Clustering
Speech-To-Text Arabic Chinese
Name-Entity Detection
Name Entity Translation & Transliteration
Machine Translation Arabic-English Chinese-English
Trans Models
UIMA Analytics
Content Manager
Search Indexes
Storage Filter
Indexer
Model Adapt
OmniFind

Text Processing Flow



Hash Table

ID Model

Token Models

N-E Models

Trans Models

Duplicate Detection

Language ID

Detag

Tokenize Text

Name-Entity Detection

Name Entity Translation & Transliteration

Machine Translation

UIMA Analytics

Content Manager

Search Indexes

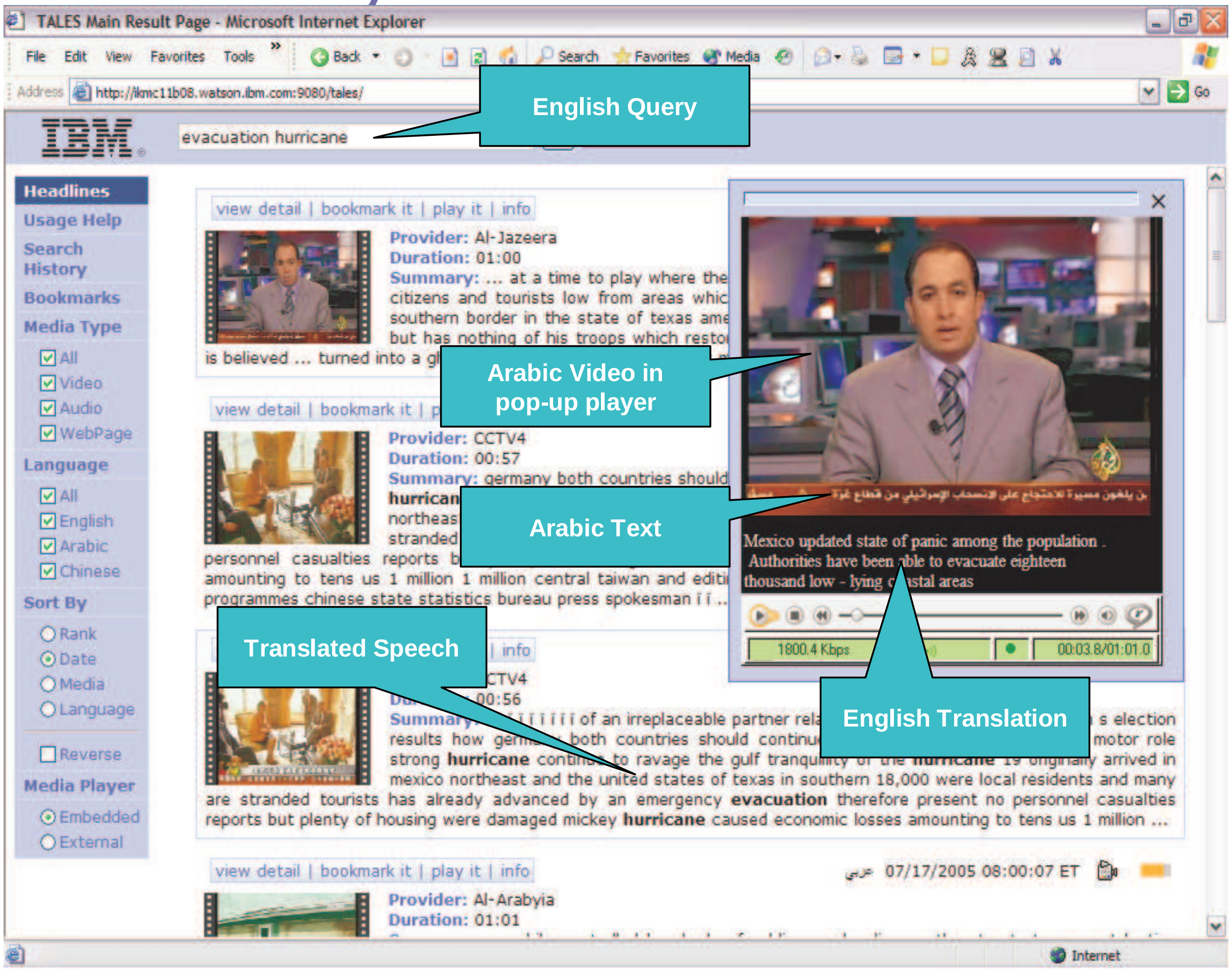
Storage Filter

Indexer

Model Adapt

OmniFind

TALES Foreign Broadcast Video Monitoring and Search System



http://ikmc11b08.watson.ibm.com:9080/tales/

evacuation hurricane

Headlines
Usage Help
Search History
Bookmarks
Media Type
☑ All
☑ Video
☑ Audio
☑ WebPage
Language
☑ All
☑ English
☑ Arabic
☑ Chinese
Sort By
○ Rank
● Date
○ Media
○ Language
□ Reverse
Media Player
● Embedded
○ External

view detail | bookmark it | play it | info

Provider: Al-Jazeera
Duration: 01:00
Summary: ... at a time to play where the citizens and tourists low from areas whic southern border in the state of texas ame but has nothing of his troops which resto is believed .... turned into a gl

view detail | bookmark it | p

Provider: CCTV4
Duration: 00:57
Summary: germany both countries should **hurrican** northeas stranded personnel casualties reports b amounting to tens us 1 million 1 million central taiwan and editi programmes chinese state statistics bureau press spokesman ï ï ...

| info

CTV4
Du 00:56
Summary ï ï ï ï ï ï of an irreplaceable partner rela s election results how ger both countries should continu motor role strong **hurricane** contin to ravage the gulf tranquility of the **hurricane** is originally arrived in mexico northeast and the united states of texas in southern 18,000 were local residents and many are stranded tourists has already advanced by an emergency **evacuation** therefore present no personnel casualties reports but plenty of housing were damaged mickey **hurricane** caused economic losses amounting to tens us 1 million ...

view detail | bookmark it | play it | info

عربي 07/17/2005 08:00:07 ET

Provider: Al-Arabyia
Duration: 01:01

Mexico updated state of panic among the population .
Authorities have been able to evacuate eighteen
thousand low - lying coastal areas

1800.4 Kbps
00:03.8/01:01.0

English Query

Arabic Video in pop-up player

Arabic Text

Translated Speech

English Translation

- 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

谢谢

شكرا