

Language Research at DARPA

Joseph Olive

DARPA/IPTO

Joseph.olive@darpa.mil

The DARPA Challenge



DARPA can develop technology that –

- ❖ Removes language barriers
- ❖ Enables soldiers to:
 - Communicate with allies, enemies, local populations
 - Understand huge amounts of information available in foreign languages
 - Decipher captured documents
 - Learn foreign languages

Human - Machine Interaction

- ❖ Historical Perspective
 - Early Greeks
 - Late 18th Century
- ❖ Scientific Approaches
 - Helmholtz
 - Electronics
- ❖ Digital Computers

Human – Computer Communication

❖ Science fiction

- HAL
- C3PO

❖ Reality

- Little knowledge
- No experience
- No deductive reasoning
- No language comprehension

Speech Recognition

Late '60s

- ❖ Small vocabulary (10 words)
- ❖ Speaker dependent
- ❖ Low accuracy (approx. 85%)

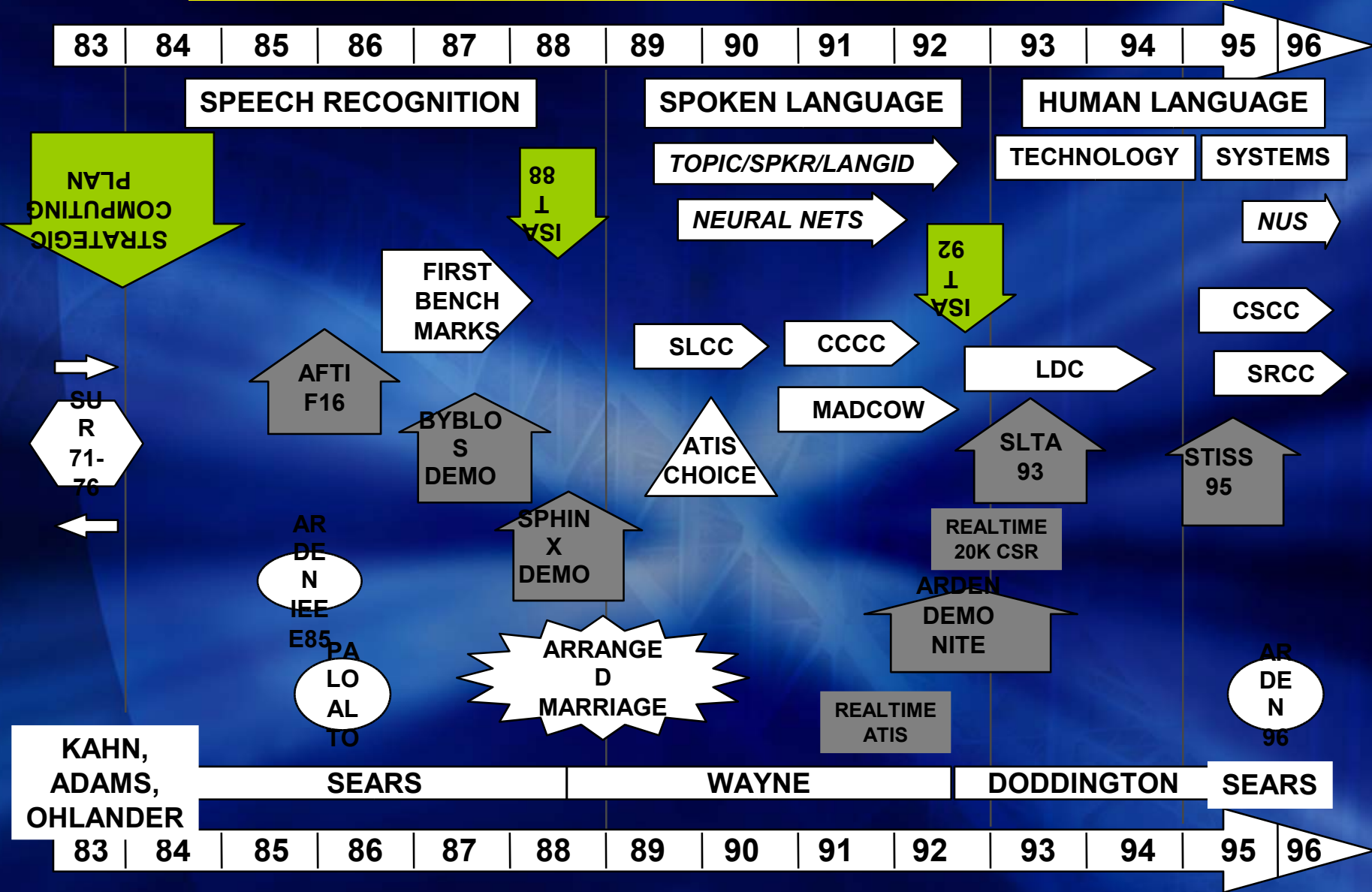
“General purpose speech recognition seems far away. Special purpose speech recognition is severely limited. It would seem appropriate for people to ask themselves why they are working in the field and what they expect to accomplish.”

-- John Pierce, 1969

Advances in Speech Recognition

- ❖ Speaker independence
- ❖ Large vocabularies
 - Word spotting
 - Concept spotting
 - Robotic operators
 - Call centers
 - Travel reservations
- ❖ Unconstrained vocabularies
 - Dictation systems
 - Closed captioning

DARPA's SPOKEN LANGUAGE PROGRAMS 1983-1996: SOME SIGNIFICANT EVENTS



Fielded Applications

❖ Phraselator – Speech to Speech Translation –

- Restricted domain
- 1 - way communication



❖ Media Monitoring Systems

➤ Features

- Unrestricted speech & text – Broadcasts & newswire
- Multiple languages – English, Chinese, Arabic
- Transcription, translation and search capabilities

➤ eTAP (enhanced Text & Audio Processor)

- First installation – May '04
- Seven installations are deployed or planned

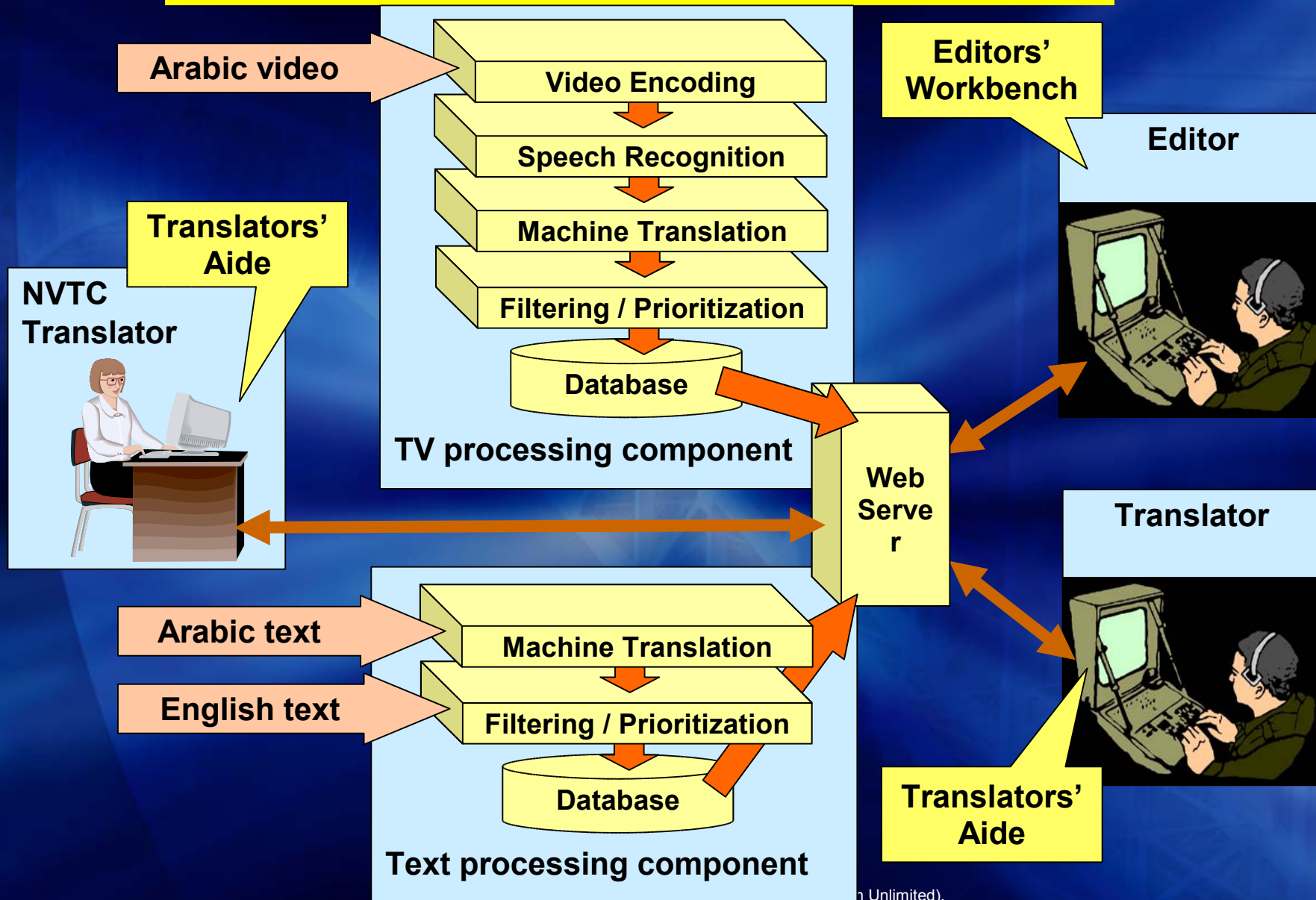


➤ Tales (Translingual Automated Language Exploitation System)

- First delivery Dec. 13, 2005
- Committed to transition to two sites



eTAP-Arabic System Components



Search over built-in sources

Built-in sources accumulate continuously

Foreign sources translated by machines

Persistent searches filter the built-in sources

Requests for human translations

Media digest being created in MS Word

Hyperlinks created automatically when document is cited

Search: allawi

Sources

- Al-Ahram: 6232 documents
- Al-Hayat: 2713 documents
- Al-Ittihad: 5447 documents
- Al-Jazeera.NET: 1048 documents
- allAfrica.com: 1953 documents
- Al-Quds Al-Arabi: 7873 documents
- Al-Sharq Al-Awsat: 10557 documents
- Al-Watan: 8513 documents
- ArabicNews.com: 361 documents
- ArabNews: 640 documents

Watchlists

Translations

- 9 translations complete
- 1 documents to be translated

Editor's Workbench Report

7 July 2004

GulfNews.com on 07 Jul 2004: [Insurgents detonate car bomb at mourning tent](#)

tonated a car bomb yesterday outside a massive tent packed with hundreds of mourners, including the victims of an earlier attack, killing 14 people and wounding dozens of others.

The blast left a 1.5 meter (yard) crater in the ground, set five cars on fire and burnt the tent canopy in the central town of Khalis, in the heart of the region.

Dismembered corpses littered the floor. White plastic chairs were overturned and twisted.

Ln 18 Col 1 REC TRK EXT OVR

The CENTCOM Experience

- ❖ Before May '04
 - Staff of 2 people
 - Open Source – English only
- ❖ May '05
 - Staff of 10 people
 - 5000 Arabic documents per week
 - 300 Sent to NVTC for human translation
- ❖ March '06
 - Staff of 19 people
 - Self sustaining operation
 - 4 Linguists
 - 24x7 operation
 - 1500 RFIs per year

What Can We Offer

❖ Technology from Two Sources

- BBN – eTAP
- IBM – TALES

❖ Technology Capabilities

- Present:
 - 55% accuracy of translation from text
 - 35% accuracy of translation from speech
 - Arabic and Chinese
 - Open source and more
 - Triage quality
 - Search capabilities
 - Work load reduction by 95%

What Can We Offer

❖ Technology Capabilities (continued)

➤ Next Year (GALE)

- 75% accuracy from text
- 65% accuracy from speech
- Distillation (automatic response to RFI) at 50% of human production

➤ Five Years

- Translation accuracy of 95%
- Distillation acceding human performance
- New languages

Global Autonomous Language Exploitation (GALE)



High Level Goal

**Enable
Automated Processes &
English Speaking Soldiers and Commanders**

To Absorb & Analyze



GALE – What is new



How it is now

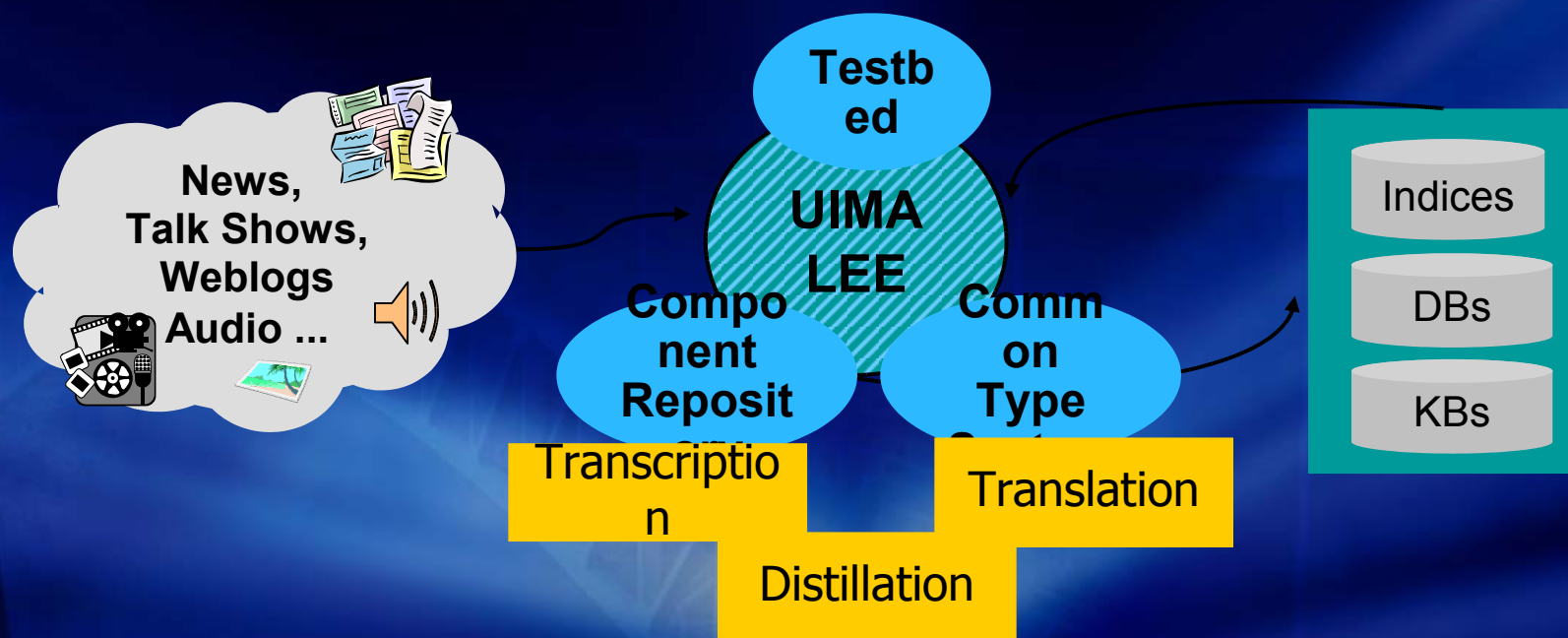
- 1. Performance advances in component technologies (EARS, TIDES)**
- 2. Statistically-based Translation with high BLEU scores**
- 3. Entity extraction and summarization**
- 4. Focus on the linguist and the analyst (J2)**
- 5. Performance driven evaluations and research**

How it will be with GALE

- 1. High performance for integrated systems in a Language Exploitation Environment**
- 2. Combined statistical AND linguistics translation with accurate meaning**
- 3. Integrated distillation of relevant information**
- 4. Focus on the operational military (J3)**
- 5. Utility AND performance driven evaluations and research**

GALE – Research Areas

- ❖ Language Exploitation Environment (LEE)
 - Integration platform
- ❖ Transcription Engine
 - Multilingual speech to English text
- ❖ Translation Engine
 - Multilingual text to English text
- ❖ Distillation Engine
 - English text to actionable intelligence
- ❖ Linguistic Data
 - Fuel for engines
- ❖ Utility Evaluation

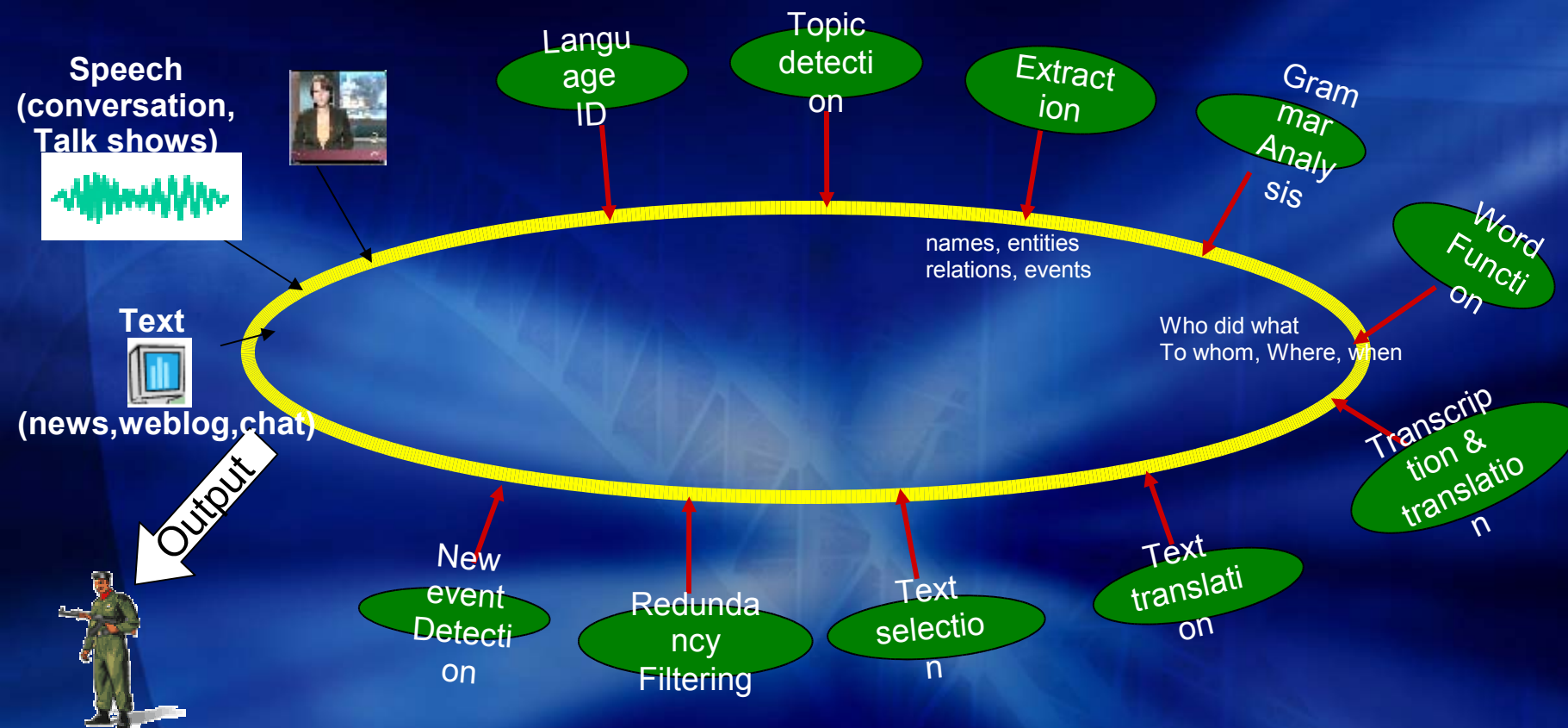


UIMA = Unstructured Information Management Architecture

- ❖ Highly-distributed plug-&-play architecture
- ❖ Software to facilitate integration & optimization
- ❖ Enables fast, scalable deployment, easy collaboration

Potential to accelerate progress in GALE & revolutionize language-based systems throughout DoD & industry

Integrated Solution

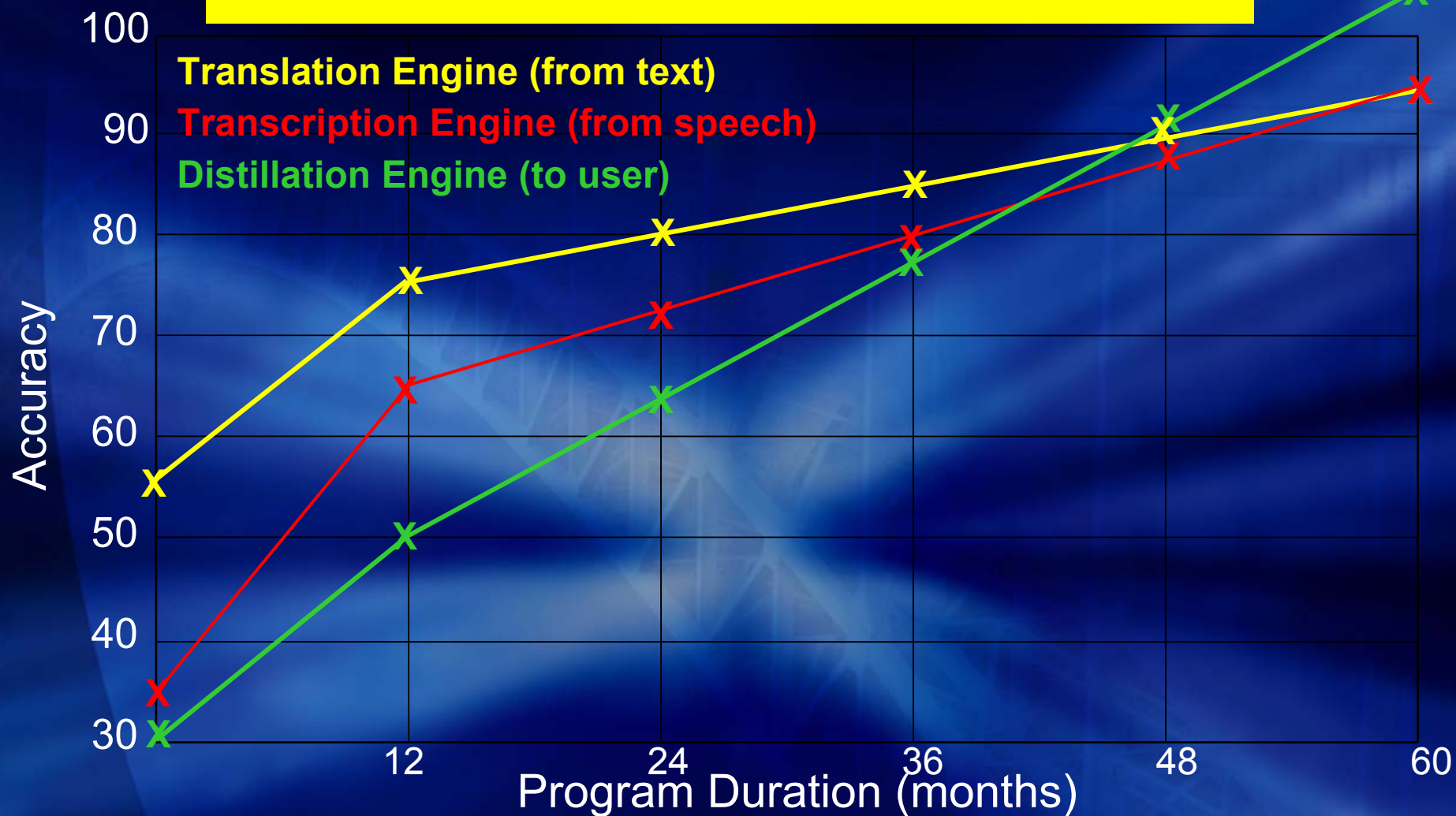


- End-to-end solution
- No intermediate solution or decision (accuracy of entire system)

Program Execution

- ❖ End-to-end technology integrated in the LEE
- ❖ Variety of important text & speech genres
- ❖ Three key languages (English, Arabic & Chinese)
- ❖ Output – English only
- ❖ End-product – Distillation
- ❖ User Interface

Minimum Targets



Accuracy: For transcription and translation – Edit distance
For distillation – Percent of human performance

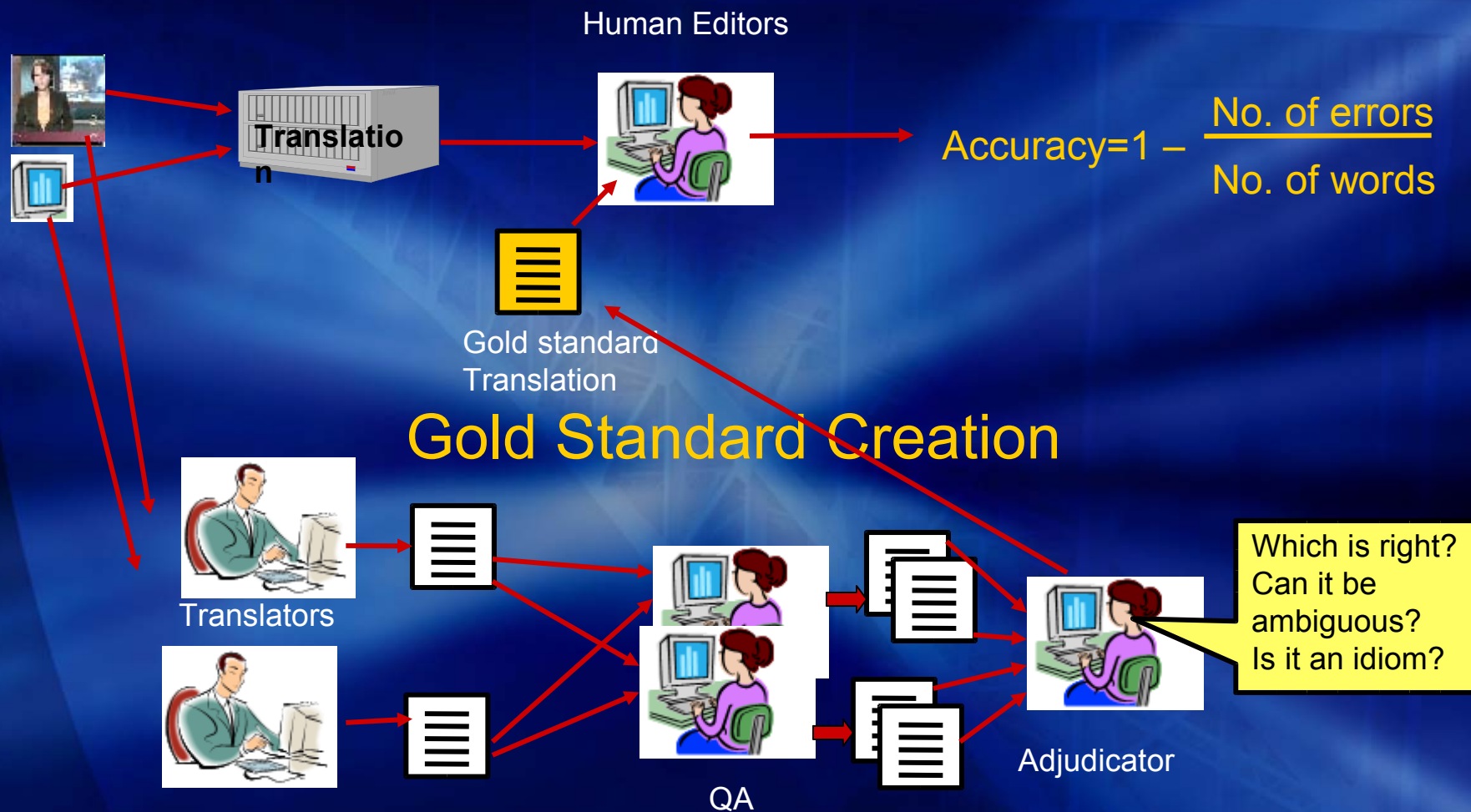
Testing Methods

- ❖ Transcription and translation
 - Machine translation is edited to reflect the meaning of human translation of the original manuscript
 - The error rate is defined as the edit distance between the original and corrected machine translation divided by the number of words in the document
- ❖ Distillation
 - Bilingual humans and machines will be presented with a set of documents in the two languages
 - They will be asked a series of questions on the documents in English
 - System input in the form of a profile or template
 - Humans will be given generic search engines and ample but finite time to accomplish the task; their answers will serve as the standards
 - Machine accuracy will be measured by subtracting irrelevant facts from the relevant ones and dividing by the total of facts listed by the human

Typical Distillation Queries

1. PRODUCE A BIOGRAPHY OF [person]
OR PROVIDE INFORMATION ABOUT [organization]
<DURING [period] ABOUT [subject]>
2. FIND STATEMENTS MADE BY OR ATTRIBUTED TO
[person] ON [topic(s)] DURING PERIOD, or IN A
LOCATION or EVENT
3. DESCRIBE THE RELATIONSHIP OF [person/org] TO
[person/org]
4. TELL ME ABOUT [person's] MEETINGS ON [topic]
5. DESCRIBE OUTBREAKS OF [disease] IN [region] IN
[time period]
6. FIND FINANCIAL TRANSACTIONS INCLUDING
MONETARY VALUE FOR [organization] IN [location]

Translation Evaluation Machine Translation



Editing Example

Example of translation accuracy measurement

Human translation:

The statement said that “your brothers in the military wing of the Al-Qaeda Jihad Organization in Mesopotamia carried out an assassination of one of the criminal tyrants in the city of Baquba.”

Corrected machine translation:

The statement said that ~~the~~ your brothers in the military wing ~~to regulate~~ of the Al Qaeda Jihad organization ~~base~~ in ~~the~~ ~~country~~ Mesopotamia had carried out the assassination of one of the criminal tyrants in the city of ~~penalty~~ Baquba.

11 errors in 33 words (67% accuracy)

Deletion
Insertion

Why new metrics

- ❖ Purpose of MT is to convey the meaning of the source
- ❖ BLEU
 - Great tool for research in the 20-50 range
 - Not sensitive to comprehensibility or accuracy
 - Favors SMT
- ❖ Other metrics
 - Meteor
 - Human assessments

New Metrics

❖ Edit distance

➤ Two versions

- Un-weighted
- Weighted

➤ Edit for preservation of meaning

❖ DLPT*

➤ Assesses content understanding

➤ Permits user inferences

➤ At 90- 95% accuracy levels edit distance better measure of quality

MT05 Sample 1

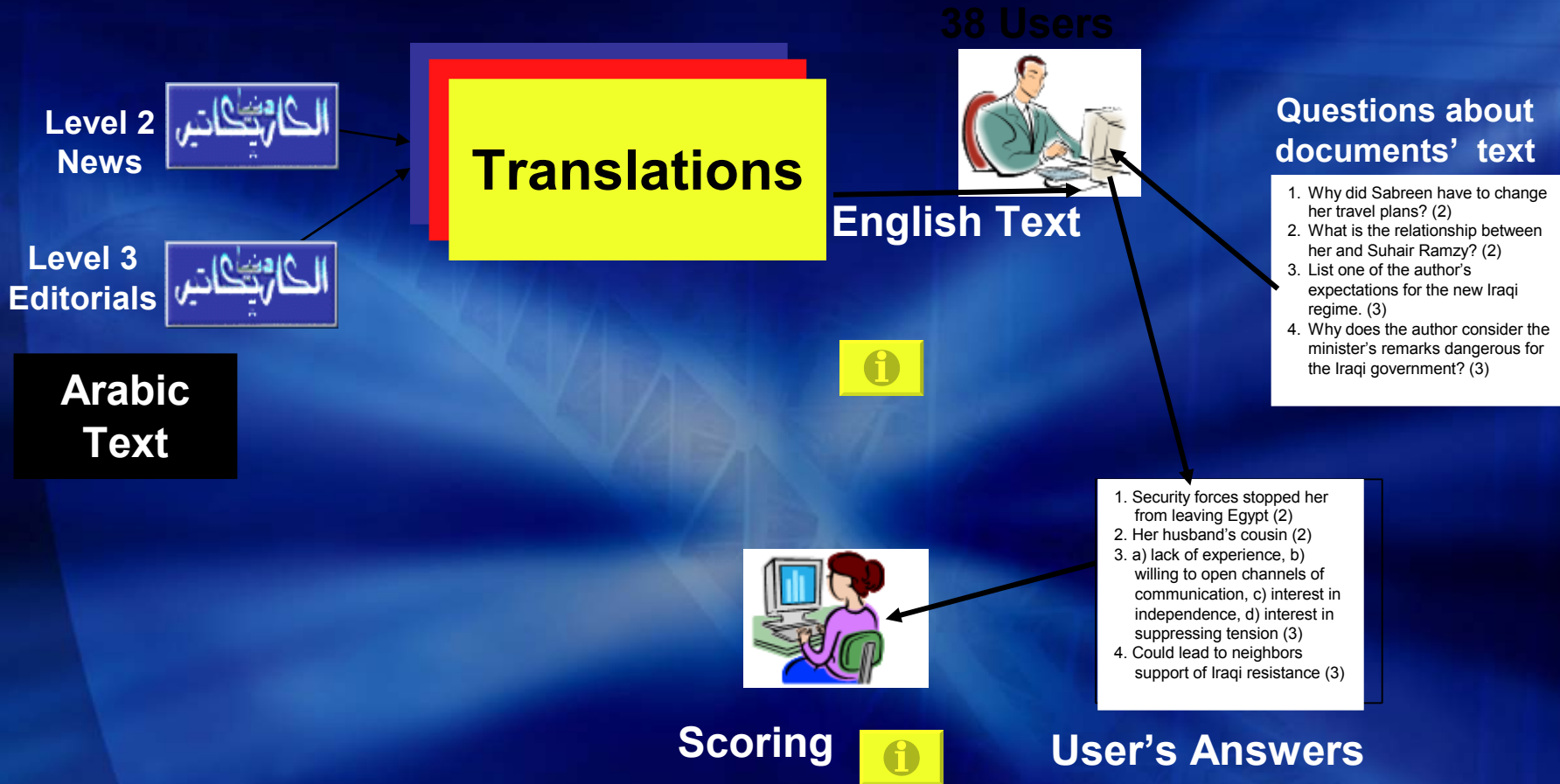
Paris 13 - 1 (AFP) - The Director General of the International Atomic Energy Agency Mohamed ElBaradei today , Monday , that the international disarmament inspectors need " a few months " to complete their mission in Iraq .

MT05 Sample 2

13 December / Xinhua / The United States denied today Monday what reports mentioned about the probability of its forces reduction in Afghanistan . _

In MT05, one system had BLEU4 score of 0.51; the other has 0.34

Which sample is from which system?



How do we get to 95%

- ❖ Rule based
 - High quality, but labor intensive
 - Difficult to improve by adding new rules
- ❖ SMT
 - Easy to implement new languages
 - Requires large corpora
 - Sensitive to domain changes
- ❖ Black box design
 - No interaction with up-stream or down stream processes


How do we get to 95%

- ❖ Probabilistic input and output
- ❖ Combining rules and statistics
- ❖ Combining different NLP statistical paradigms
 - Parsing
 - Morphology
 - Equivalent words and phrases
 - ACE type extraction – especially names
 - Analysis of who did what to whom, where, when, etc.
- ❖ New methods for Language Modeling (LM)
 - Distance relations
 - Compelling N-grams
 - Topic sensitive LM

Back-ups




[view detail](#) | [bookmark it](#) | [play it](#) | [info](#)



Provider: Al-Jazeera
Duration: 01:00
Summary: ... at a time to play where the citizens and tourists flow from areas which southern border in the state of texas arrived but has nothing of his troops which restorance is believed ... turned into a ghost town sees no where non cars ma

[view detail](#) | [bookmark it](#) | [play it](#) | [info](#)



Provider: CCTV4
Duration: 00:57
Summary: germany both countries should hurricane continue to ravage the gulf (j) northeast and the united states of texas stranded tourists has already advanced personnel casualties reports but plenty of housing were damaged amounting to tens us 1 million 1 million central taiwan and ed

Level 2 HT – Text



Cairo: “The Middle East.” Egyptian security forces at Cairo Airport prevented the retired actress Sabreen from traveling to the Kingdom of Saudi Arabia with her husband Tarek Galal Mahmoud. Sabreen was trying to travel to Saudi Arabia the day before yesterday with her husband when Egyptian authorities discovered that the actress, whose full name is Sabreen Yaseen Mahmoud Abdullah and who resides in Old Cairo's Elmanial quarter, is among those barred from leaving the country. The newspaper Elshark Elawsat was not able to elicit any response from the actress. The paper tried contacting her several times on her cell phone, which seemed to have been turned off, as well as on her home phone, but she was unavailable. The actress Sabreen retired about four years ago after her last successful series, “OM Kolthoom,” without providing any reasons for her retirement. A few months ago, she got married for the second time. Her current husband, Tarek Galal Mahmoud, is the cousin of retired actress Soheir Ramzy. Sabreen has one son from her previous marriage with businessman Yaser Abdulatif.

Sample Translation Questions

Question:

2. Why did Sabreen have to change her travel plans?
3. What reason did Sabreen give for retiring from TV show business?
4. What is the family relationship between her and Suhair Ramzy?

Answer:

7. Security Forces stopped her from leaving Egypt.
8. No reason given.
9. Her husband's cousin

Credit:

12. One credit.
13. One credit.
14. One credit.