

Natural Language Processing: Challenges, Solutions, and Applications

Bonnie Dorr

May, 2013

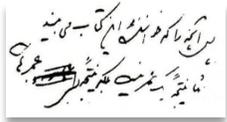
Themes

- Language understanding, translation, and summarization, require more than “just statistics”
 - Moving from high resolution (low noise) media to unrestricted and degraded or noisy media
- Linguistically-motivated approaches can benefit from the robustness of statistical/ML techniques
 - Moving from problems with general characteristics to problems applicable to real-world data.



DARPA's Language Research Programs

Hardcopy Foreign Documents



Foreign Speech



conversation

Digital Foreign Text

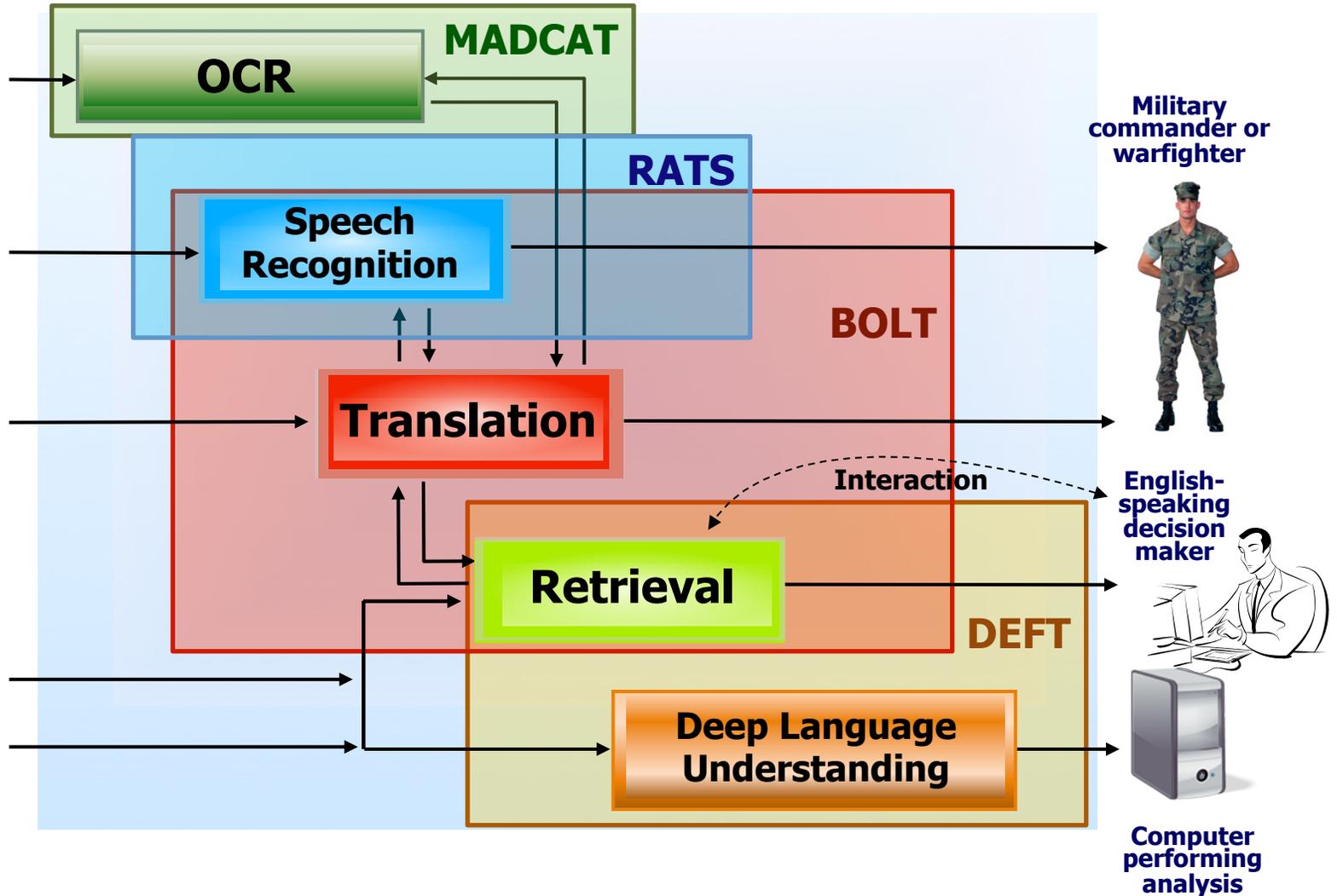


SMS, email

Digital English Text



Digital Foreign Text





What are the challenges to language research?

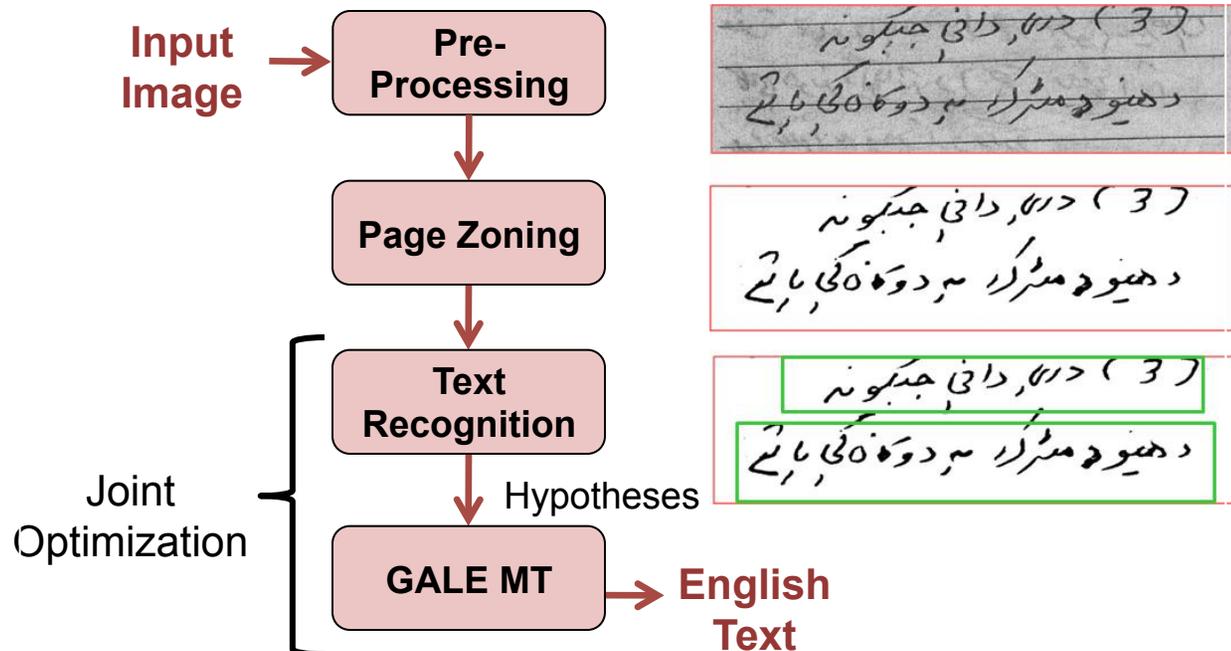
- Linguistic variety
 - Dialects
 - Cross-language divergences
- Technological shifts in forms of communication
 - Casual, verbal communication
 - Grammatical correctness in structure has disappeared.
 - Global societal change
- Volume of information
 - Astronomical increases in volume
 - Demands on human translators vastly exceed resources
 - Far surpasses human assessment capability

Automated Foreign Language Exploitation is the Key



Multilingual Automatic Document Classification, Analysis, and Translation (MADCAT)

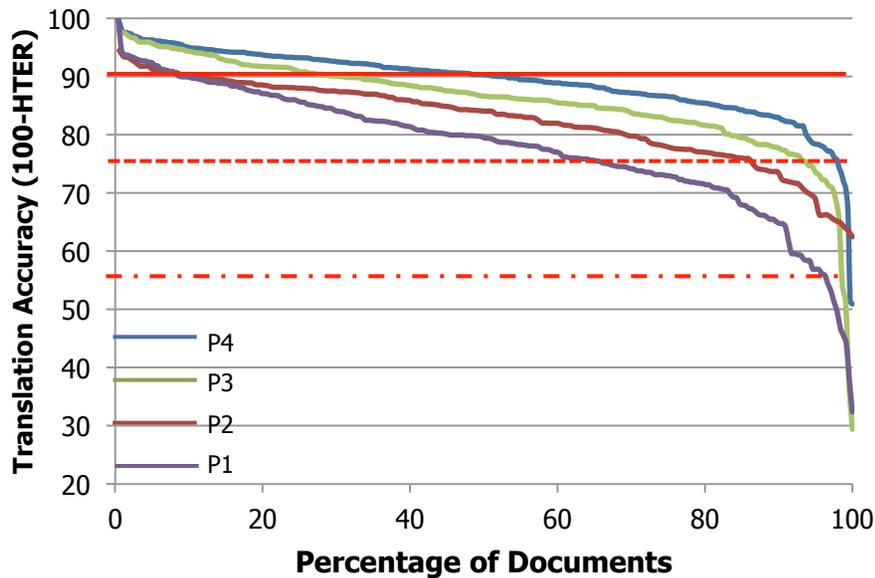
- Objective: Extract actionable info from foreign language text images
- Technical Approach:
 - Detect and recognize text in images, extract relevant metadata, and translate recognized text into English
 - Joint optimization of MADCAT component technologies and GALE machine translation (MT)



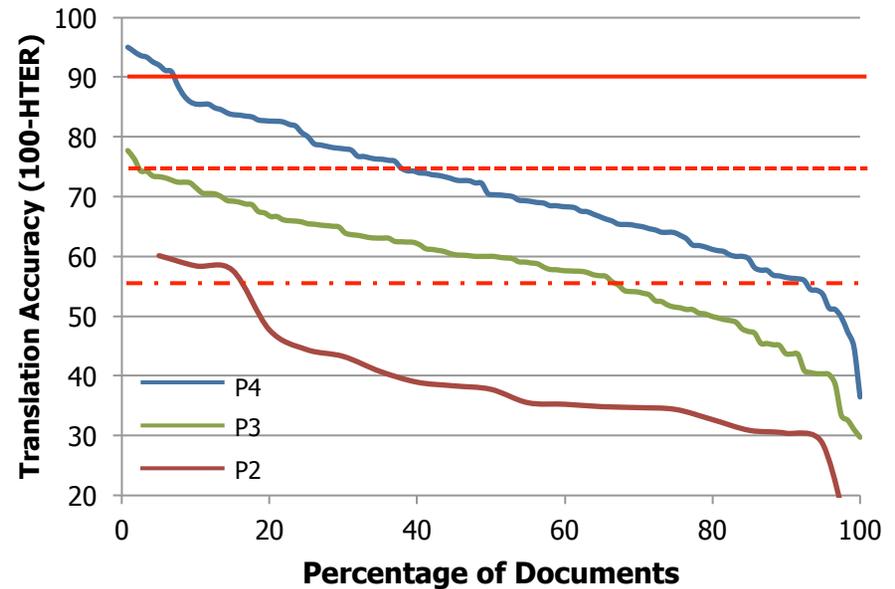


MADCAT Technical Progress

MADCAT-Generated Data



Data collected in Iraq in 1991

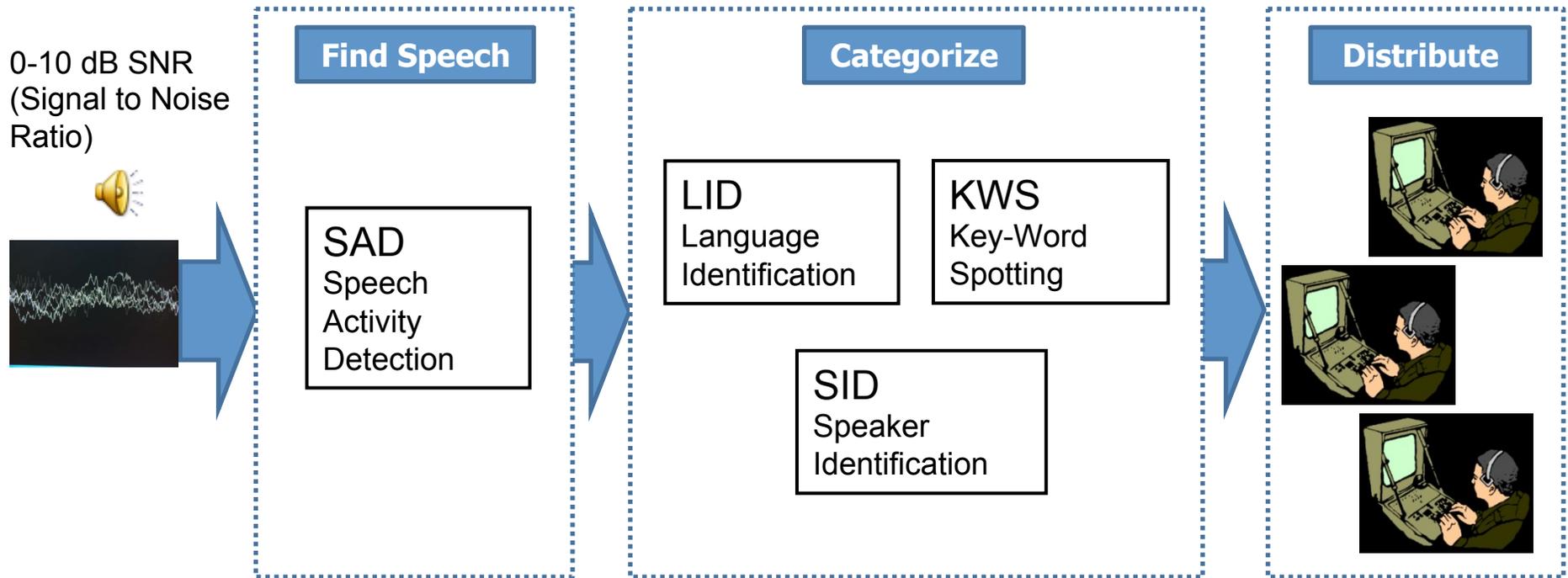


- Editable
- - - Gistable
- · · Triageable



Robust Automatic Transcription (RATS)

Goal: Create technologies for exploitation of potentially speech-containing signals received over extremely noisy and/or distorted communication channels



Improve Capability to Find and Make use of Foreign Language Speech Signals



Broad Operational Language Translation (BOLT)

Program Goal: Informal Language and Multi-Turn Conversations

BOLT is developing natural language processing capability to enable:

1. Translation and information retrieval for informal language and
2. Bilingual, multi-turn informal conversation using text or speech

Informal language is characterized by:

- Use of dialects
- Sloppy or garbled speech or text
- Incomplete and ungrammatical sentences
- Frequent references by use of pronouns
- Frequent changes in topic
- Interjection of disfluencies (restarts, interjections - “uh” and fill words – “you know”)

Baseline MT System Error Rate for Arabic → English Text

Formality	Material Type	Dialect	Accuracy
Formal	newswire	MSA	95%
Semi formal	blogs & news groups	MSA	87%
Semi formal	various web media	Dialectal Arabic	67%
Informal*	messaging	Dialectal Arabic	<40%



BOLT Target Applications: Examples

Handling Dialects

Handling dialects is crucial for automated processing of informal Arabic web material

Arabic Variant	Arabic Source Text	Pre-BOLT MT
Modern Standard Arabic	لايوجد كهرباء، ماذا حدث؟	Does not have electricity, what happened?
Egyptian Regional Dialect	الكهرباء اتقطع، ليه كده بس؟	Atqtat electrical wires, Why are Posted?
Levantine Regional Dialect	شكلكو مفيش كهرباء، ليه ينش هي كده؟	Cklo Mafeesh كهرباء, Lech heck?
Reference:	There is no electricity, what happened?	

Iraqi Regional Dialect	شو م الكو كهرباء، خير؟ ت يقوتل اع تيق وتل اع صب تيوتي ر لم عت ام لب ق لوول	Xu MACON electricity, good?
Reference:	before you retweet, check the Time lol	
Pre-BOLT MT:	T. Ikuatl AZ AZ Tel Tik casting Tioti t not signed or the core of S to Wall	



BOLT Target Applications: Chinese Examples

Pronoun and null subject/object resolution

大家心里都能猜出几分，越是这样控制舆论越让人们群众心里起疑。

Reference: Everyone can guess in their hearts, and the more they try to control public opinion this way, the more the people become suspicious in their hearts.

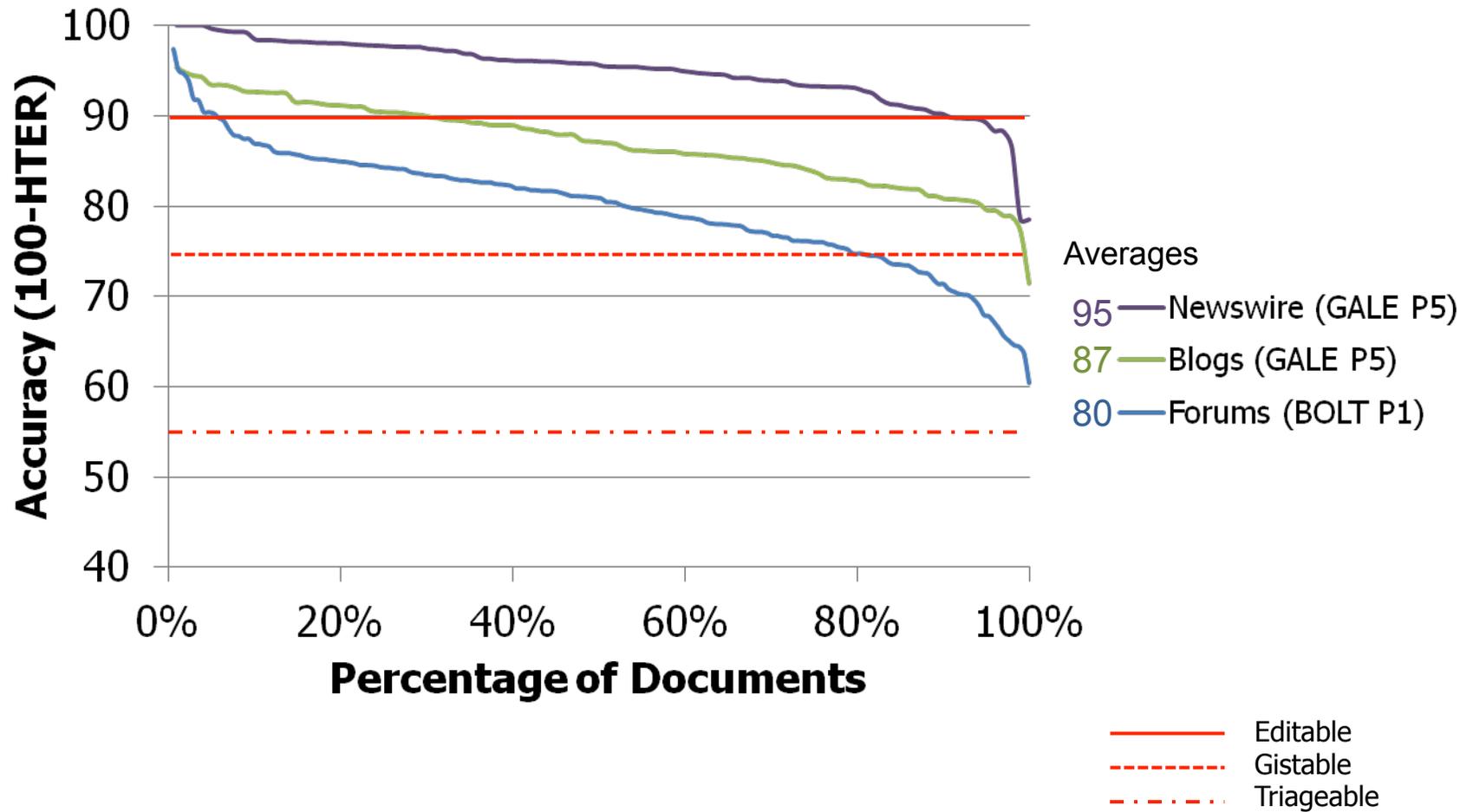
Literal: Everyone heart in can guess some, more is thus control public opinion more cause masses heart in arise suspicion.

Pre-BOLT MT: We all know can guess a bit, the more people that control the mass media more suspicious mind.



BOLT Evaluation Results - Machine Translation

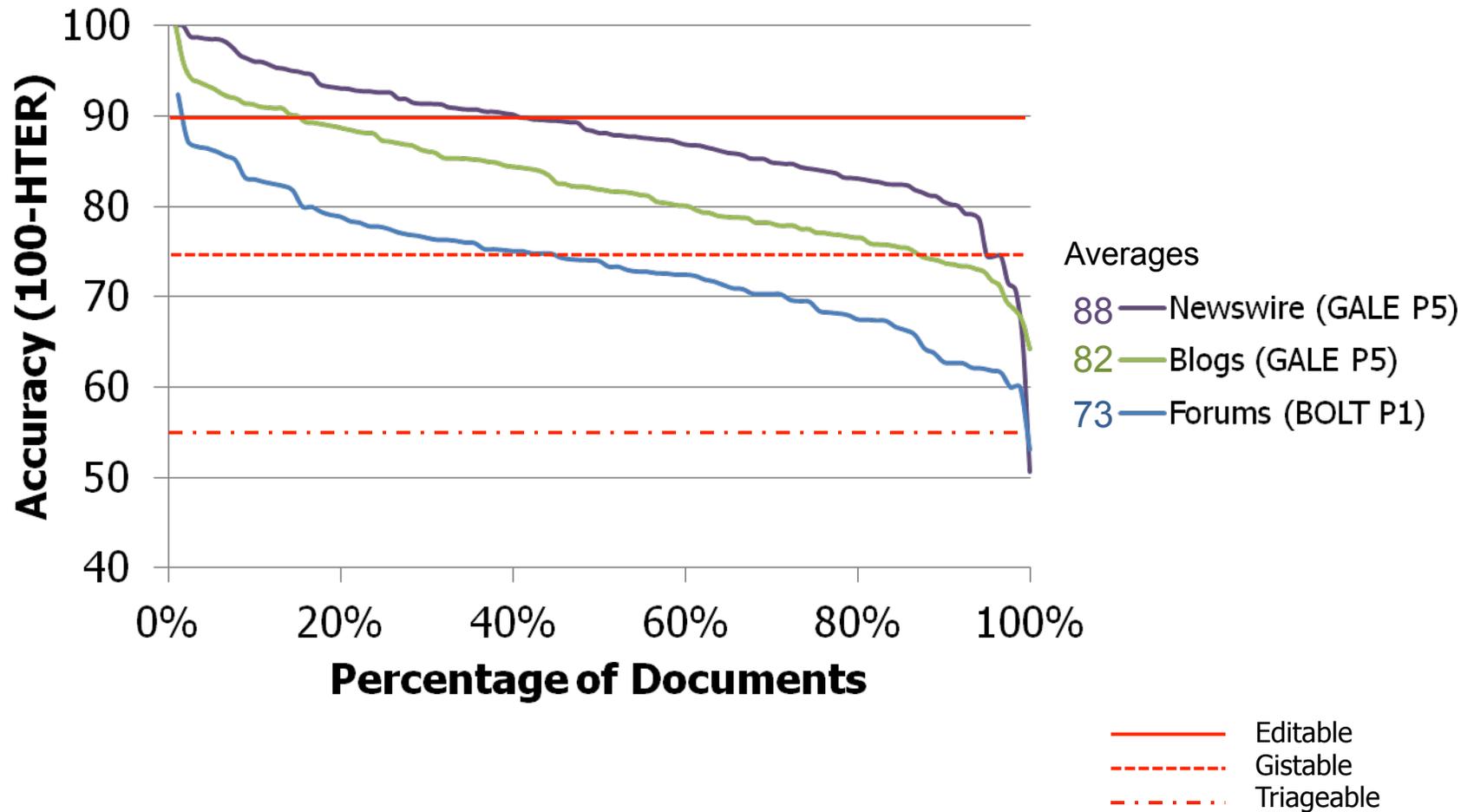
Arabic to English Text Translation





Evaluation Results - Machine Translation

Chinese to English Text Translation





Clarification Dialogue Example

Script:	The < wines > will not be allowed on the sites.
ASR of User:	The wind this will not be allowed on the sites 
Clarification1	Please give me a different word or phrase for this: (USER SPEECH) 
ASR of User:	the alcohol 
Clarification2	Did you mean allowed as in permitted, or aloud as in using one's voice? Please say one or two 
ASR of User:	one 
Transcript	the alcohol will not be allowed on the sites
Translation	الكحول ما رح يصير بالمواقع



Clarification Dialogue – Before and After

BOLT Demonstration for Out of Vocabulary Names

Approved for Public Release, Distribution Unlimited

The MT Challenge: Linguistic Divergences

- Expressing the underlying concept of a set of words in one language using a different structure in another language
- Experiments indicate that these occur in 1/3 of sentences in certain language pairs (e.g., English-Spanish).
- Proper handling of linguistic divergences:
 - enriches translation mappings for statistical extraction
 - improves the quality of word alignment for statistical MT.

Ah-hah! Back to our theme.

Divergence Categories

- Light Verb Construction
To butter → poner mantequilla (put butter)
- Manner Conflation
To float → ir flotando (go floating)
- Head Swapping
Swim across → atravesar nadando (cross swimming)
- Thematic Divergence
I like grapes → me gustan uvas (to-me please grapes)
- Categorical Divergence
To be hungry → tener hambre (have hunger)
- Structural Divergence
To enter the house → entrar en la casa (enter in the house)



Generation-Heavy Hybrid MT (GHMT)

- Motivating Question: Can we inject statistical techniques into linguistically motivated MT?
- Using “approximate Interlingua” for MT
 - Tap into richness of deep target-language resources
 - Linguistic Verb Database (LVD)
 - CatVar database (CATVAR)
- Constrained overgeneration
 - Generate multiple linguistically-motivated sentences
 - Statistically pare down results

[Work with Nizar Habash and Christof Monz, 2009]

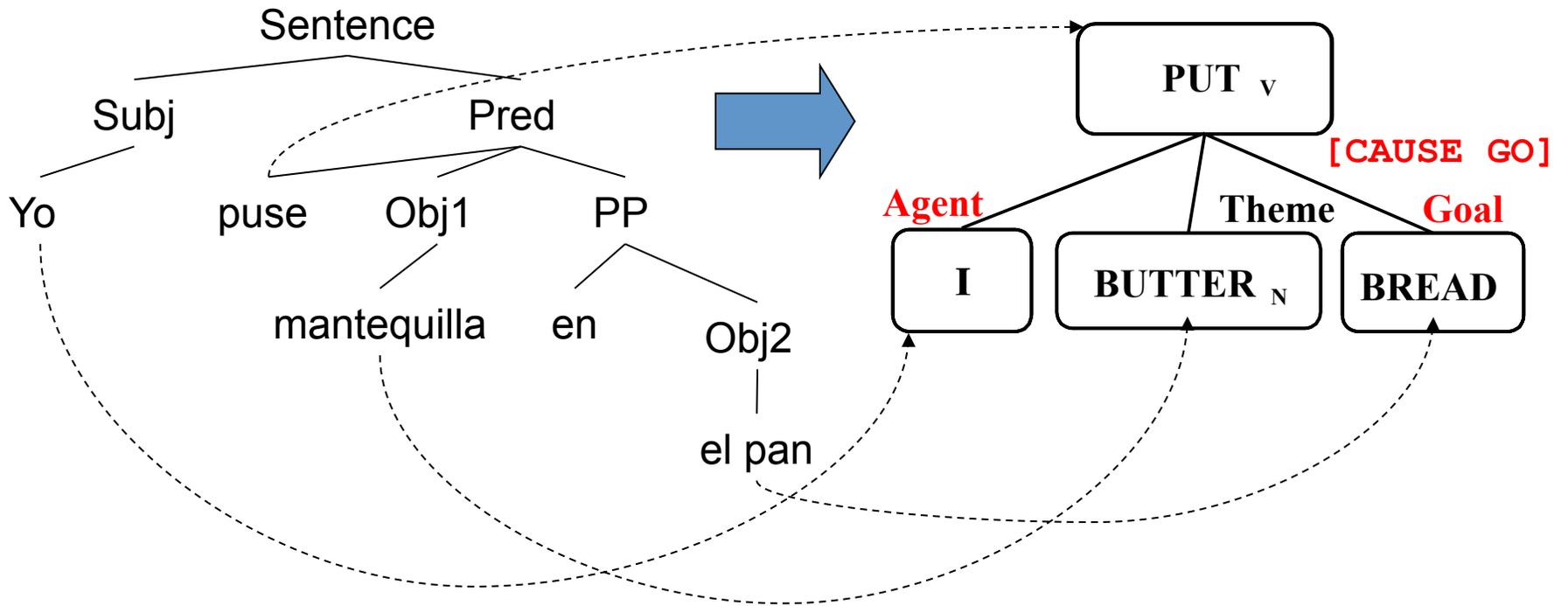
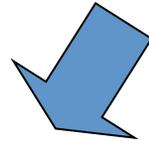


UNIVERSITY OF
MARYLAND

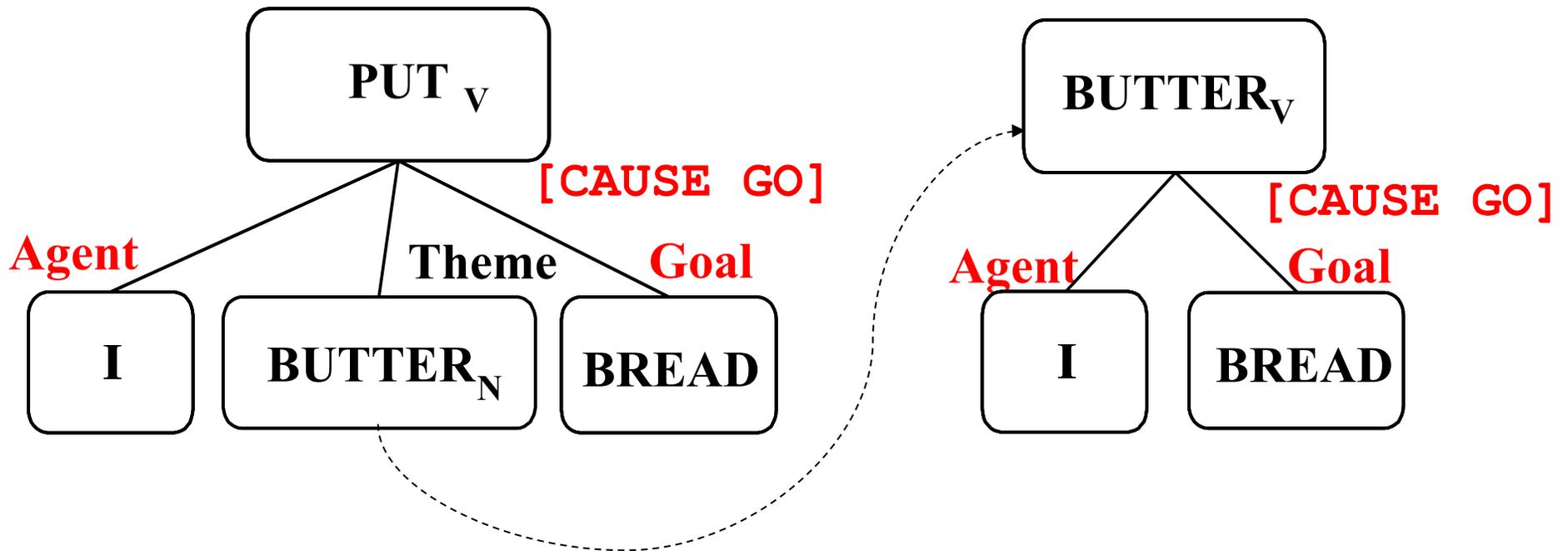


GHMT Example

Yo puse mantequilla en el pan



GHMT Example (continued)



Knowledge Resources in English only
(LVD; CATVAR - Dorr, Habash, Monz, 2003, 2006, 2009)

GHMT: Statistical Extraction after Linguistic Generation (Language Model induced from ML)

X **puse mantequilla en** Y \longrightarrow X **buttered** Y
(*X put butter on Y*)

Rank	Hypothesis
1	I buttered the bread
2	I butter the bread
3	I breaded the butter
4	I bread the butter
5	I buttered the loaf
6	I butter the loaf
7	I put the butter on bread



MT System Combination Findings

- Combination of approaches (“Hybrid MT” and “Linguistically informed Stats MT”) achieves better MT results than either approach alone (Habash & Dorr, 2006)
- Best paper award, NAACL 2007: "Combining Outputs from Multiple Machine Translation Systems" (Ayan&Dorr @ University of Maryland and Rosti&Schwartz @ BBN)
- Hybrid approaches are now the standard for large-scale MT systems.
- Jacob Devlin, former UMD student, exploring richer combination approaches. (Best paper, NAACL-2012)



UNIVERSITY OF
MARYLAND



What types of linguistic knowledge
(beyond syntactic parsing) are useful?



Linguistic Annotations and Inferred Knowledge Units

First Order Knowledge	Second Order Knowledge
Modality: X __ happen can <permissive> must <obligative> might <possible>	Confidence value regarding the occurrence of event X
Dialog Acts: Send me Y <request-action> Did Y get sent <request info> I sent Y <inform>	Inferred relationships and intentions Requestee is subordinate of requester "send Y" is an intended activity



Inferring relationships and intentions from informal communication

- Patterns of interaction reflect social situation (who has power, who has status, etc) [Passonneau & Rambow, 2009]
- Patterns of interaction (taken together with modality/confidence) reveal implicit relationships and underlying intentions [Discussions with IHMC 2011]
- Use of Modality and Negation for Semantically Informed Machine Translation [Dorr et al., 2012 *Computational Linguistics*, 38:2]
- Opinion Analysis for detecting *Intensity*, not just positive vs. negative. [U.S. Patent 8,296,168, October 23, 2012, with Subrahmanian, Reforgiato, and Sagoff].





Deep Exploration and Filtering of Text (DEFT)

Example Information Communication

A: Where were you? We waited all day for you and you never came.

B: I couldn't make it through, there was no way. They...they were everywhere. Not even a mouse could have gotten through.

A: You should have found a way. You know we need the stuff for the...the party tomorrow. We need a new place to meet...tonight. How about the...uh...uh...the house? You know, the one where we met last time.

B: You mean your uncle's house?

A: Yes, the same as last time. Don't forget anything. We need all of the stuff. I already paid you, so you had better deliver. You had better not \$%*! this up again.

Co-referring Locations

Inter-related Events

Anomaly, Novelty, Emerging trends

Causal Relations (why, how)



What would DEFT produce?

Informal Communication Input

A: Where were you? We waited all day for you and you never came.

B: I couldn't make it through, there was no way. They...they were everywhere. Not even a mouse could have gotten through.

A: You should have found a way. You know we need the stuff for the...the party tomorrow. We need a new place to meet...tonight. How about the...uh...uh...the house? You know, the one where we met last time.

B: You mean your uncle's house?

A: Yes, the same as last time. Don't forget anything. We need all of the stuff. I already paid you, so you had better deliver. You had better not \$%*! this up again.

DEFT Summary Bullets

- **People**
 - Person A – Superior; planner.
 - Person B – Inferior; courier.
 - Unspecified group “we”
- **Associations**
 - B has previously received a call from a person of interest
- **Activities**
 - Meeting between A and B
 - One failed attempt today
 - Planning another attempt for tonight. Time unspecified.
 - Planned delivery of “stuff” by B to A
- **Causal Links**
 - Attempted meeting did not happen due to security in area of meeting.
- **Geospatio-Temporal Links**
 - The planned event is going to happen tomorrow
 - Location: “the house”, point of call origination
- **Entity-Event Linking**
 - Replanned meeting associated with “your uncle's house”
 - Previous meetings at “your uncle's house”
 - B paid an unspecified amount for services.

Information from other Sources

The Future

- Global shift to new forms of communication
 - Informal communication, dialects, and implicitly conveyed info
- Focus on problems and data with real-world applicability
 - Express technical progress in terms understandable to end users (e.g., editable, gistable, triageable)
- Generation-Heavy Hybrid MT
 - More robust across genre
 - System combination produces best results
- Inferring Relations and Intentions from informal communication
 - Requires deeper linguistic knowledge
 - Potential for hybrid linguistic/statistical approach to inference

Questions?

bonnie.dorr@darpa.mil

