# Machine Translation Challenges, Solutions, and Applications

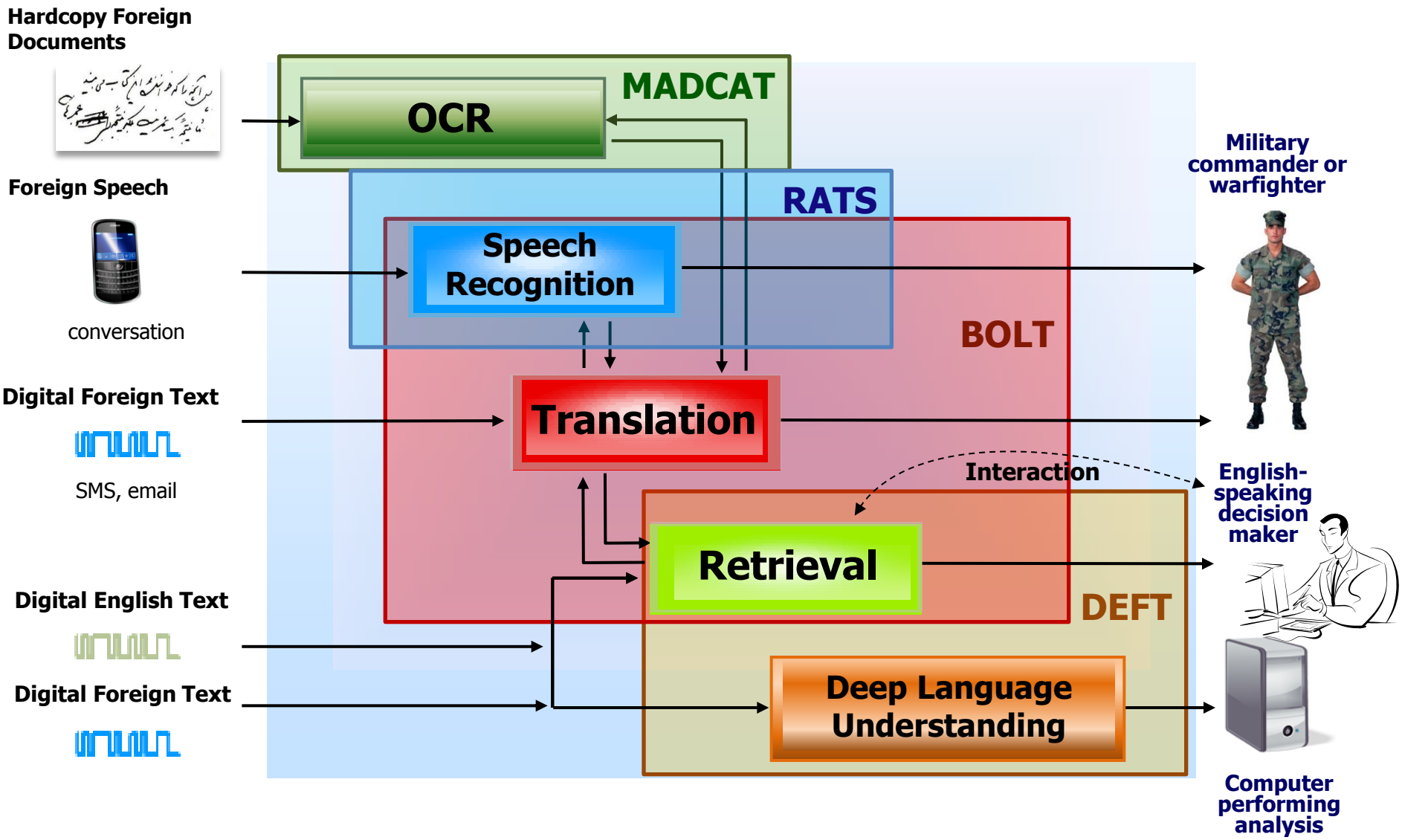
Bonnie Dorr

11 September 2013

# Themes in MT Research and Applications

- 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




Dorr, AMTA-2012 Keynote Speech, October, 2012

# Challenges to Machine Translation Research

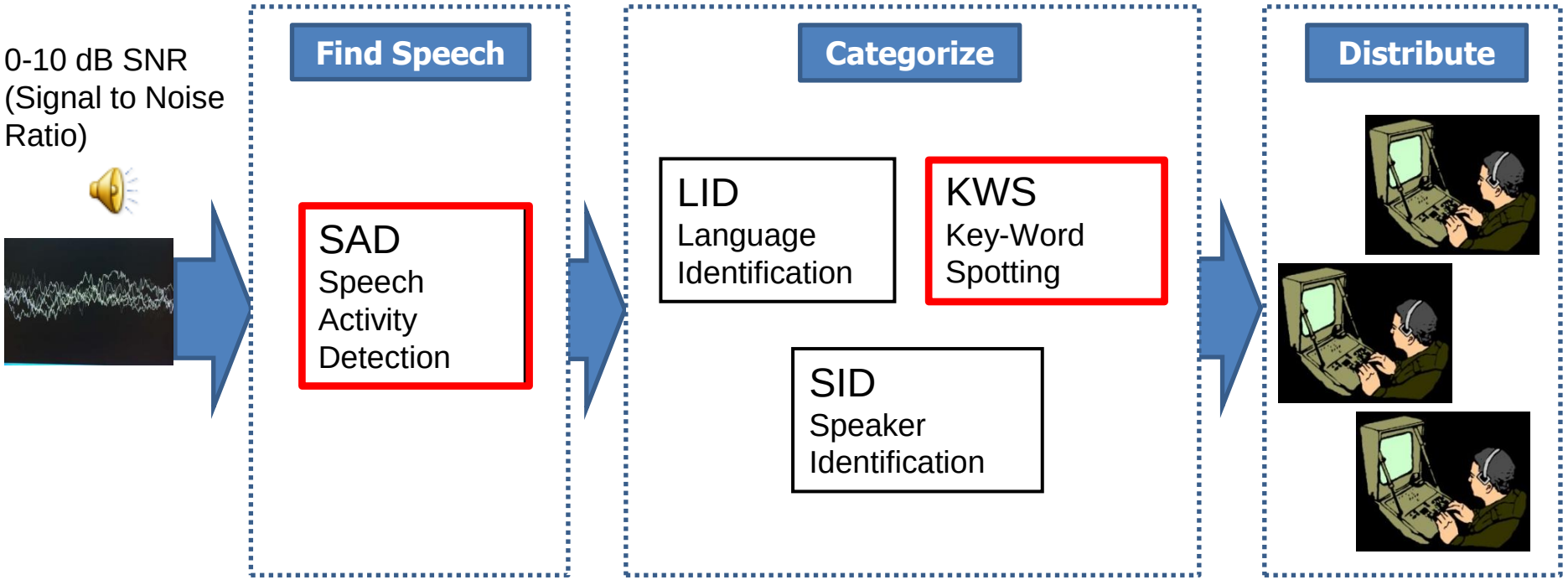
- Linguistic variety
  - Dialects
  - Cross-language divergences
- Technological shifts in forms of communication
  - Casual, verbal communication
  - Grammatical correctness in structure has disappeared.
  - Global societal change
- Volume of information
  - Astronomical increases in volume
  - Demands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

**Automated Foreign Language Exploitation is the Key**

Dorr, AMTA-2012 Keynote Speech, October, 2012

# Robust Automatic Transcription (RATS)

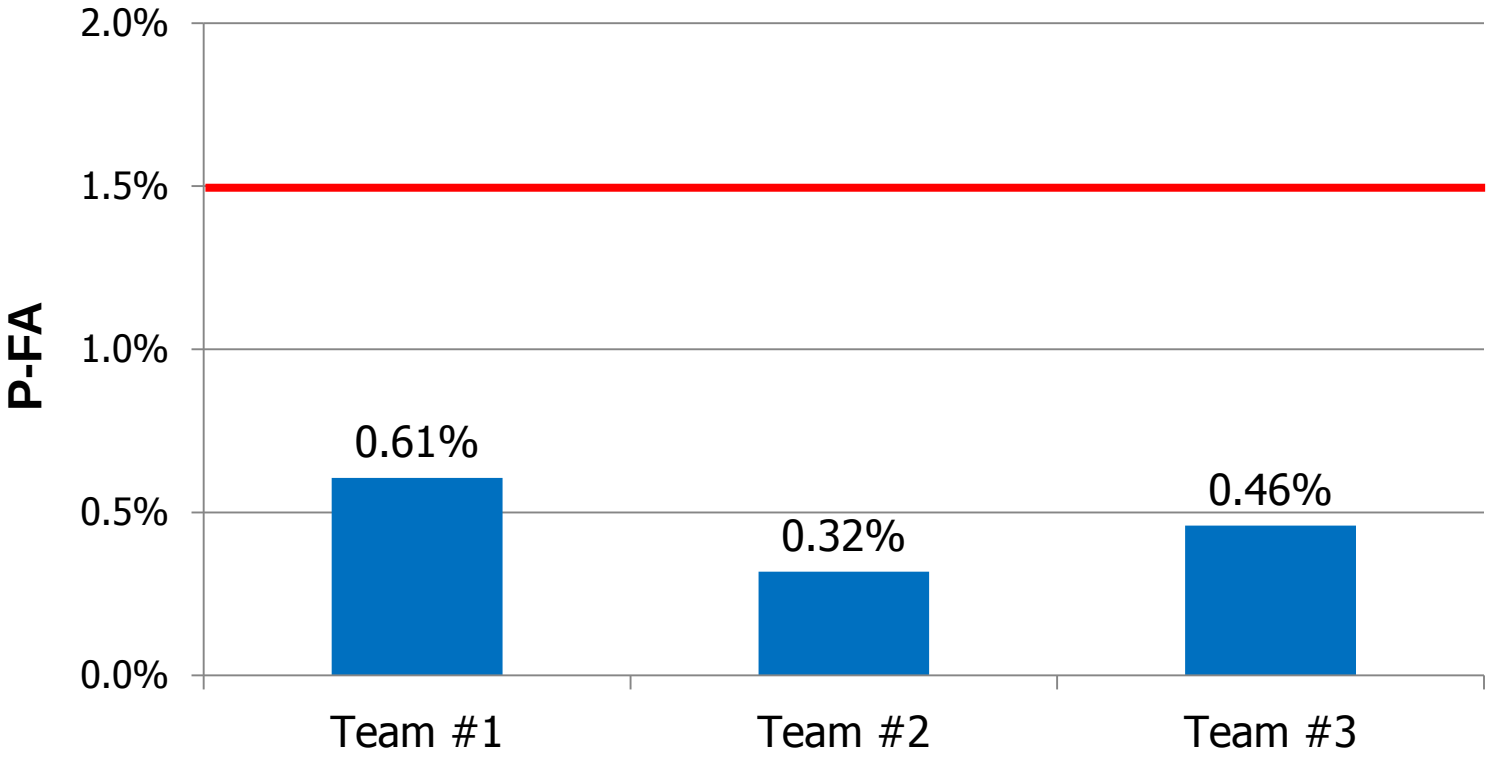
Goal: Create technologies for exploitation of potentially speech-containing signals received over extremely noisy and/or distorted communication channels




**Improve Capability to Find and Make Use of Foreign Language Speech Signals**

Dorr, AMTA-2012 Keynote Speech, October, 2012

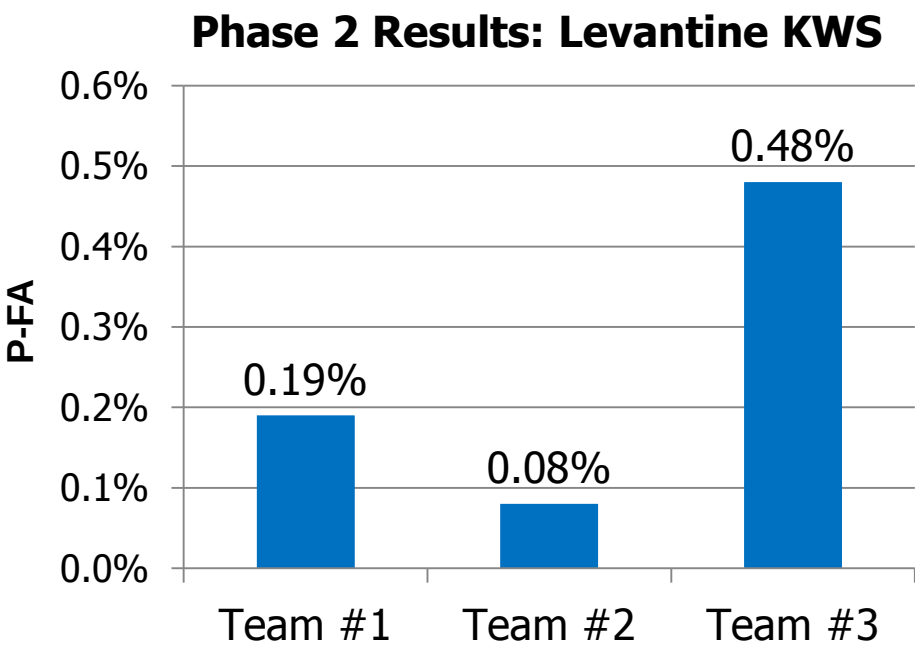
# SAD False Alarm Rates at 4% Pmiss (Phase 2 Target)



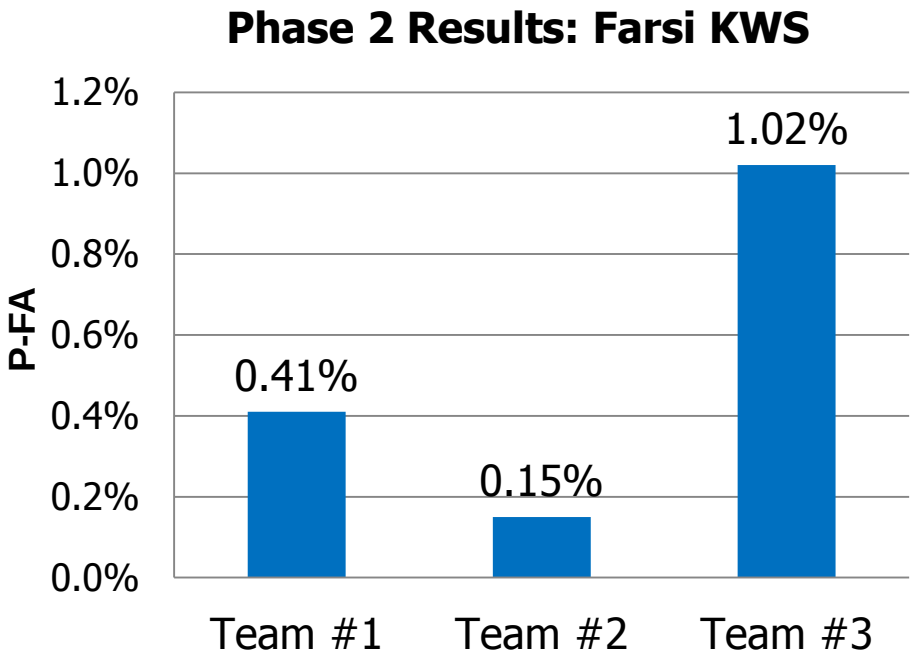

- Results for SAD shown in terms of probability of false alarm at target probability of misses (4%)
- All teams met the target for phase 2

# RATS SAD Demo (SRI)

# KWS Levantine and Farsi False Alarm Rates at P2 Miss Rate (20%)

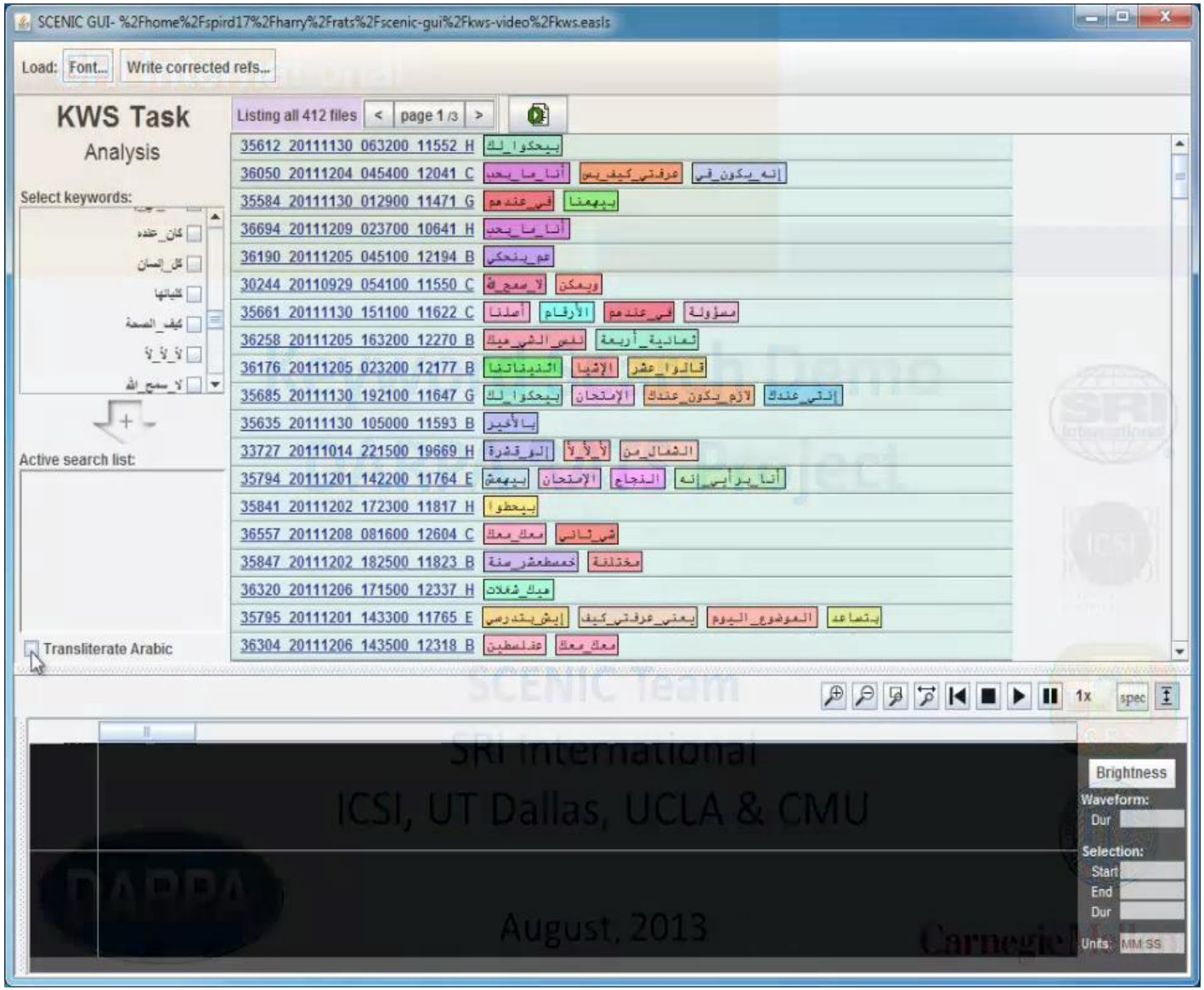



Phase 2 Target off the chart
(3% False Alarm at 20% Miss)



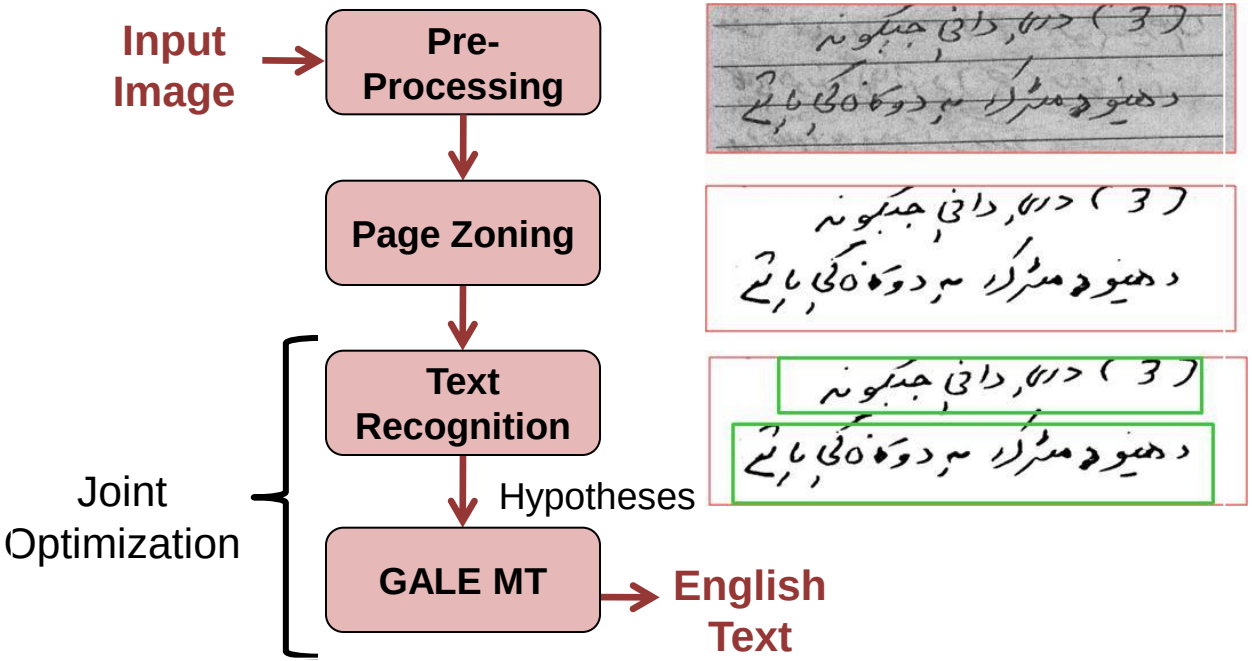

Phase 2 Target off the chart
(3% False Alarm at 20% Miss)

# RATS KWS Demo (SRI)



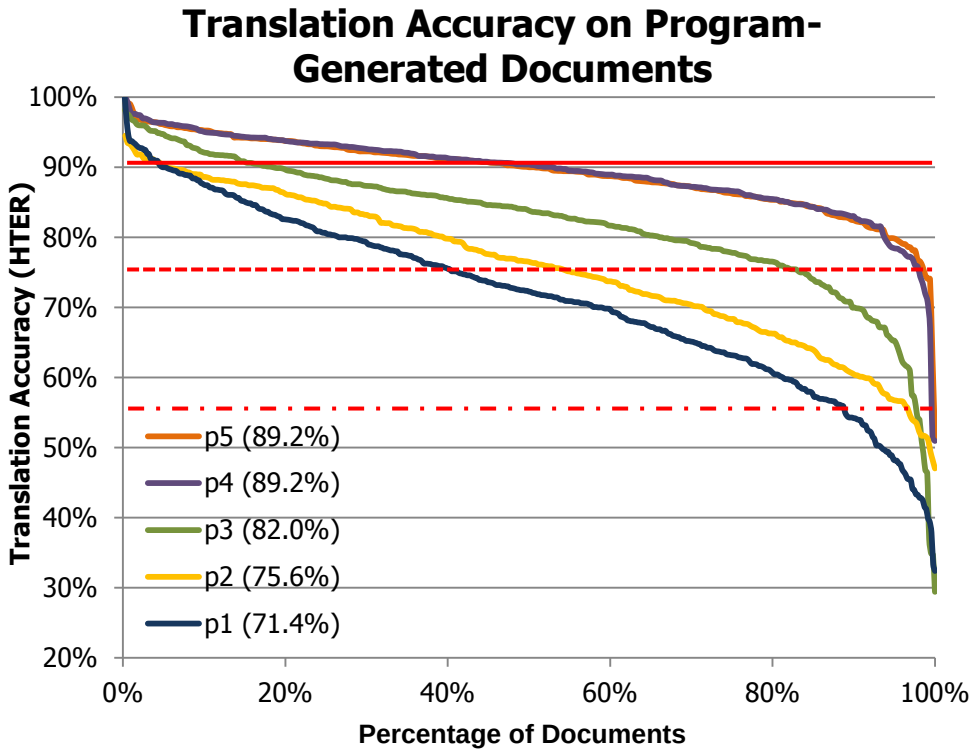
# Multilingual Automatic Document Classification, Analysis, and Translation (MADCAT)

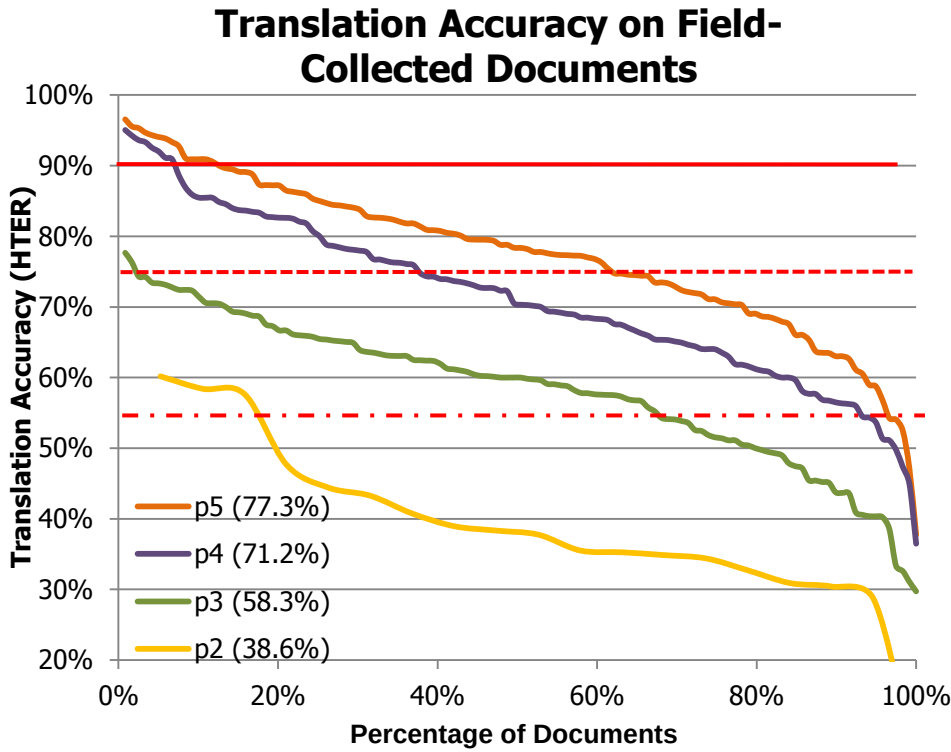
- Objective: Extract actionable info from foreign language text images
- Technical Approach:
  - Detect and recognize text in images, extract relevant metadata, and translate recognized text into English
  - Joint optimization of MADCAT component technologies and GALE machine translation (MT)




Dorr, AMTA-2012 Keynote Speech, October, 2012

# Translation Results




Editable
Gistable
Triageable

# 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% |

Dorr, AMTA-2012 Keynote Speech, October, 2012

*Estimated degradation

# 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 | الكهربا اتقطعت، ليه كده بس؟ | Atqtat electrical wires, Why are Posted? |
| Levantine Regional Dialect | شكلو مفيش كهربا، ليش هيك؟ | Cklo Mafeesh كهربا, Lech heck? |
| Iraqi Regional Dialect | شو ماكو كهرباء، خير؟ | Xu MACON electricity, good? |

Reference: There is no electricity, what happened?

ت يقوتل ا ع تيق وتل ا ع صب تيوتي ر لم عت ام لب ق لوول

Reference: before you retweet, check the Time lol

Pre-BOLT MT: T. Ikutl AZ AZ Tel Tik casting Tioti t not signed or the core of S to Wall

# BOLT Target Applications: Chinese Examples

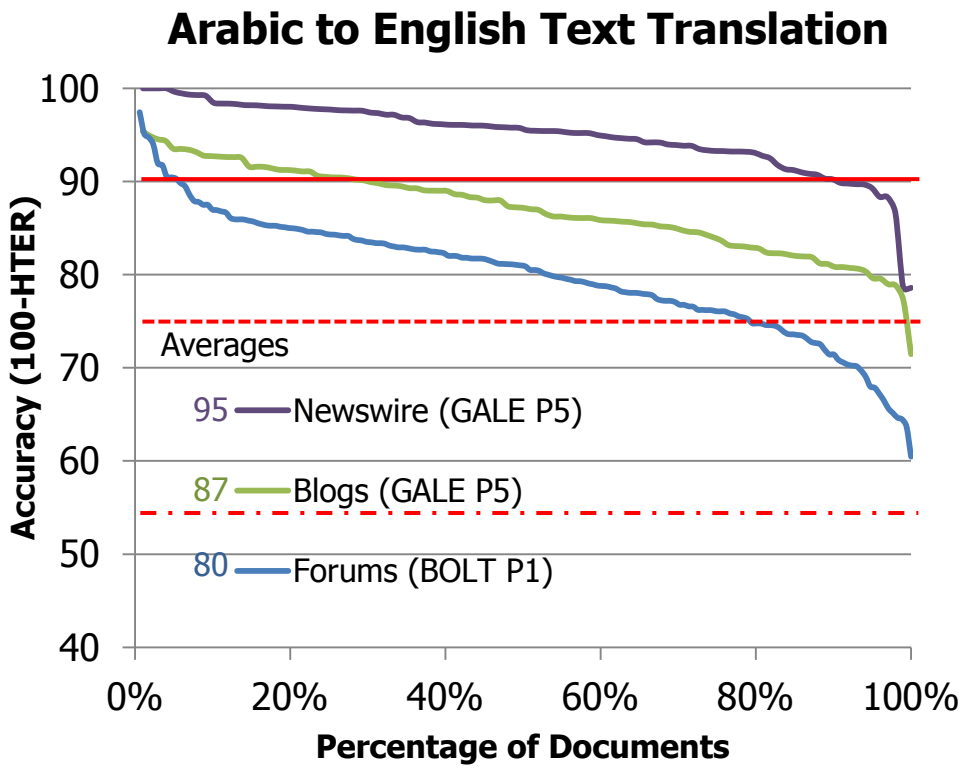
## Pronoun and null subject/object resolution

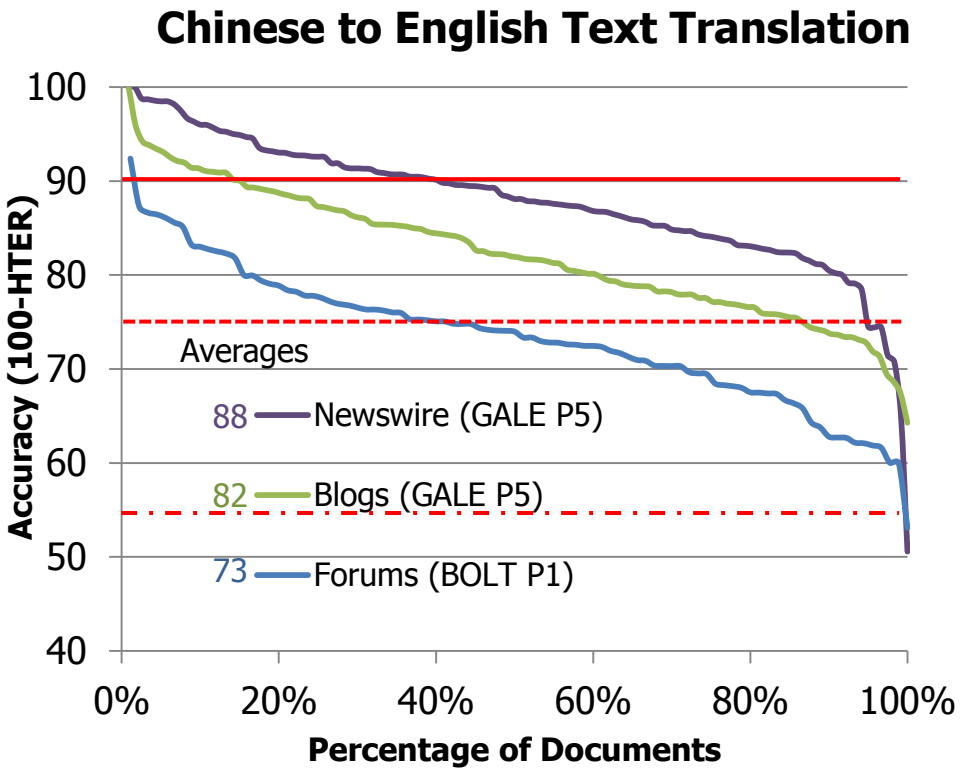
大家心里都能猜出几分，越是这样控制舆论越让人们群众心里起疑。

| | |
|---|---|
| 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. |

Dorr, AMTA-2012 Keynote Speech, October, 2012

# Machine Translation Evaluation Results




Editable
Gistable
Triageable

Dorr, AMTA-2012 Keynote Speech, October, 2012

# BOLT Machine Translation Illustration (BBN)

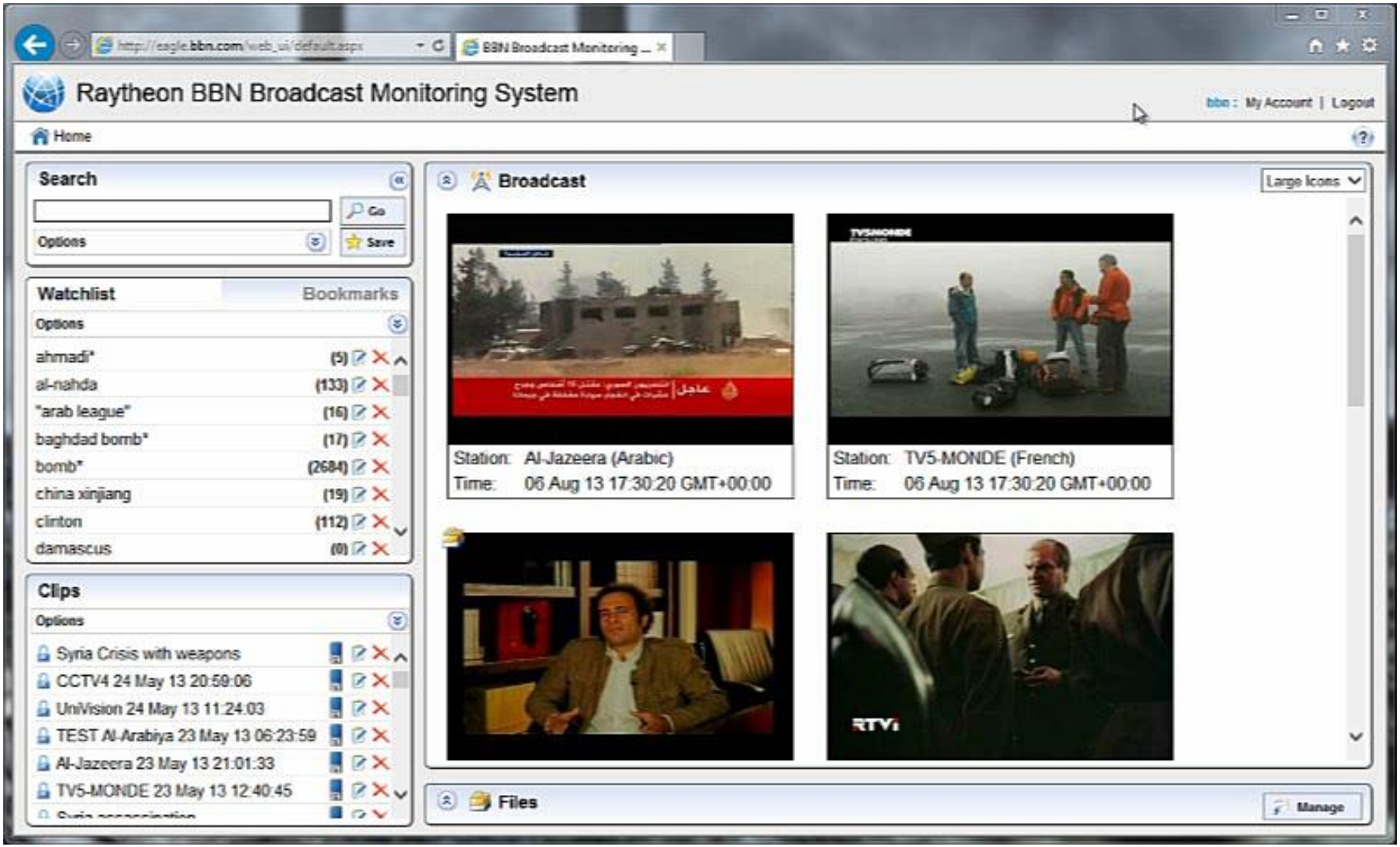
الجيش لا يعرف دستووووووووور ولايهتم باعلاااااااااااام

| | |
|---|---|
| Generic MT | The army does not knows Dstwooowoowoowor but ايهتمBaalaaaaaaaaaaaaaaam |
| BOLT MT | The army does not know the constitution and care about the media |
| Ref | The Army does not know a constitution and does not care about the media |

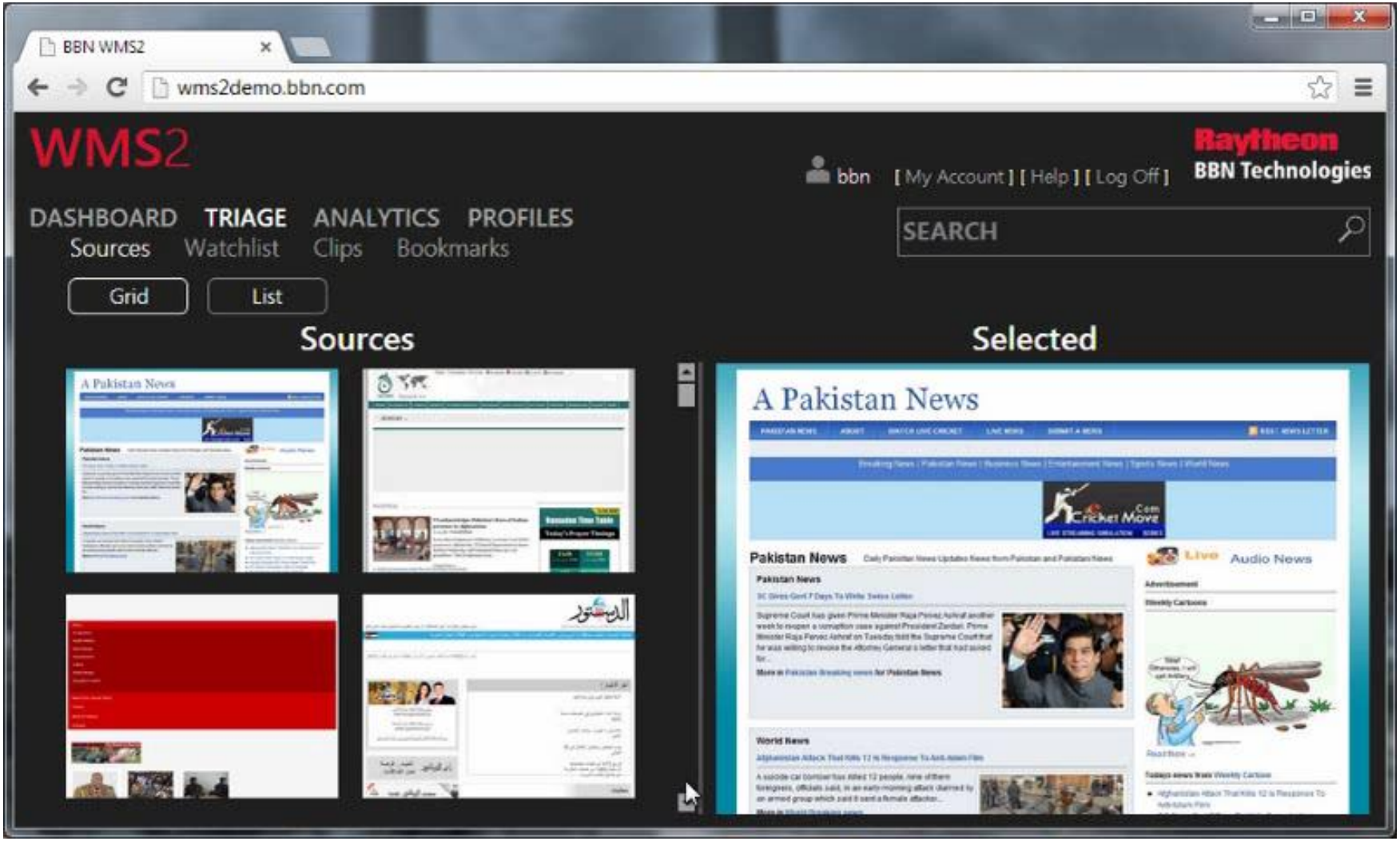
انا عشت فى شرم و عارف إن القوات اللى هناك مش قوات أمريكيه

| | |
|---|---|
| Generic MT | I lived in Sharm El-Aref The Elly forces there mesh U.S. troops |
| BOLT MT | I lived in Sharm and I know that the forces there are not American forces |
| Ref | I lived in Sharm El-Sheikh and I know that the forces there are not American forces. |

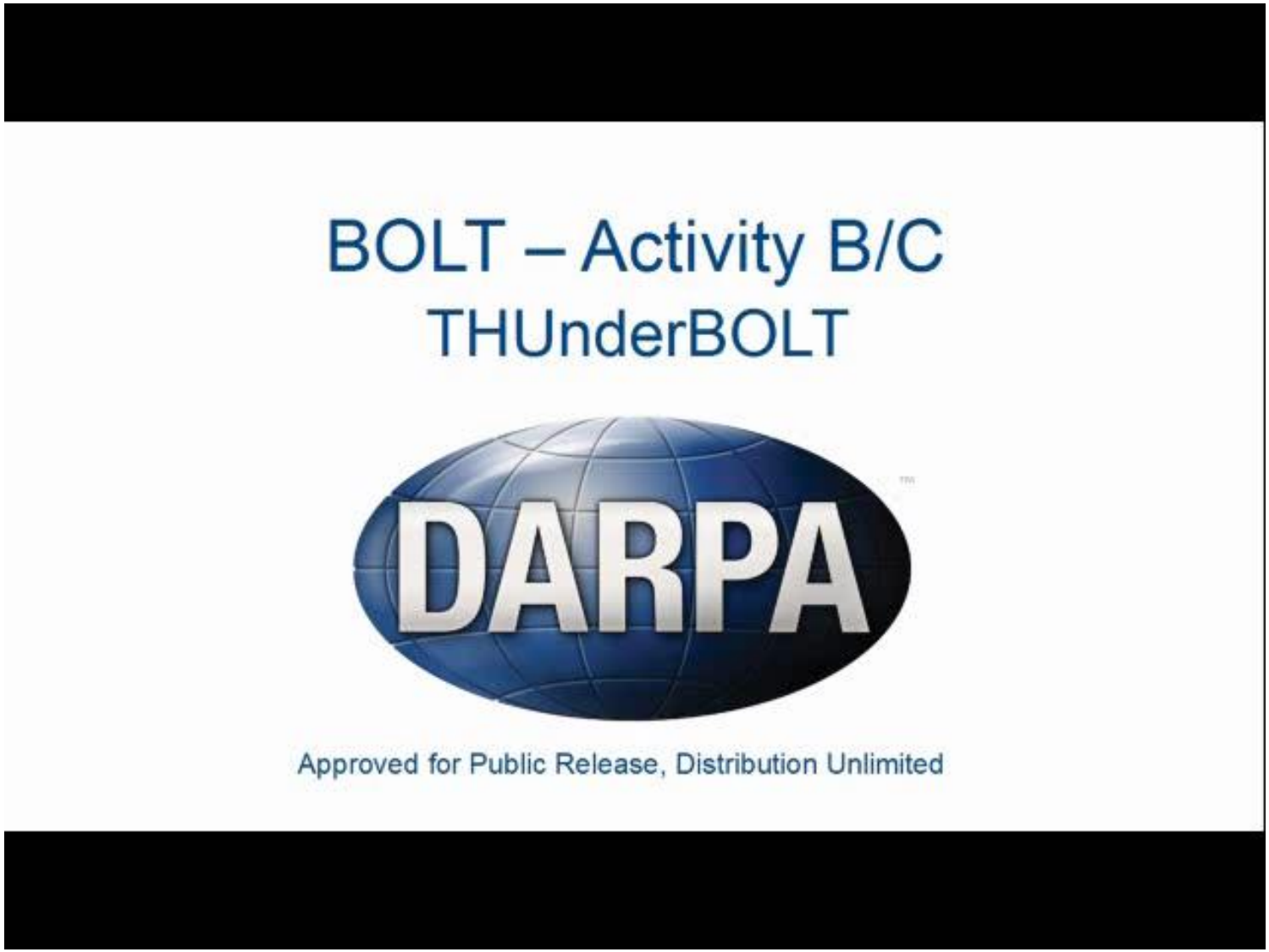
# Broadcast Monitoring System (BBN)



# Web Monitoring System (BBN)



# BOLT Speech-to-Speech Demo (SRI)



# 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

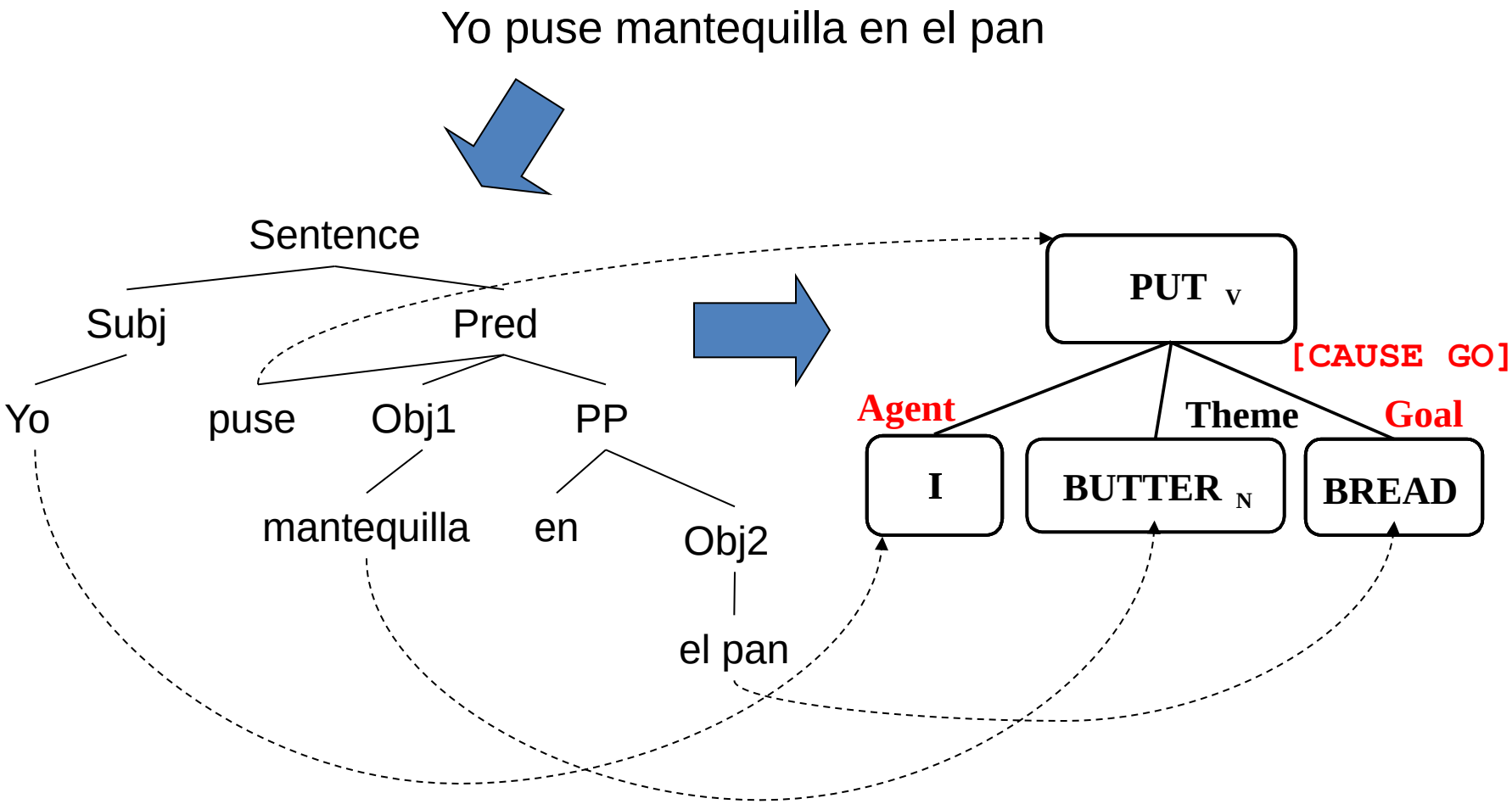
- Light Verb Construction
  - To butter → poner mantequilla (put butter)
- Manner Conflation
  - To float → ir flotando (go floating)
- Head Swapping
  - Swim across → atravesar nadando (cross swimming)
- Thematic Divergence
  - I like grapes → me gustan uvas (to-me please grapes)
- Categorial Divergence
  - 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)

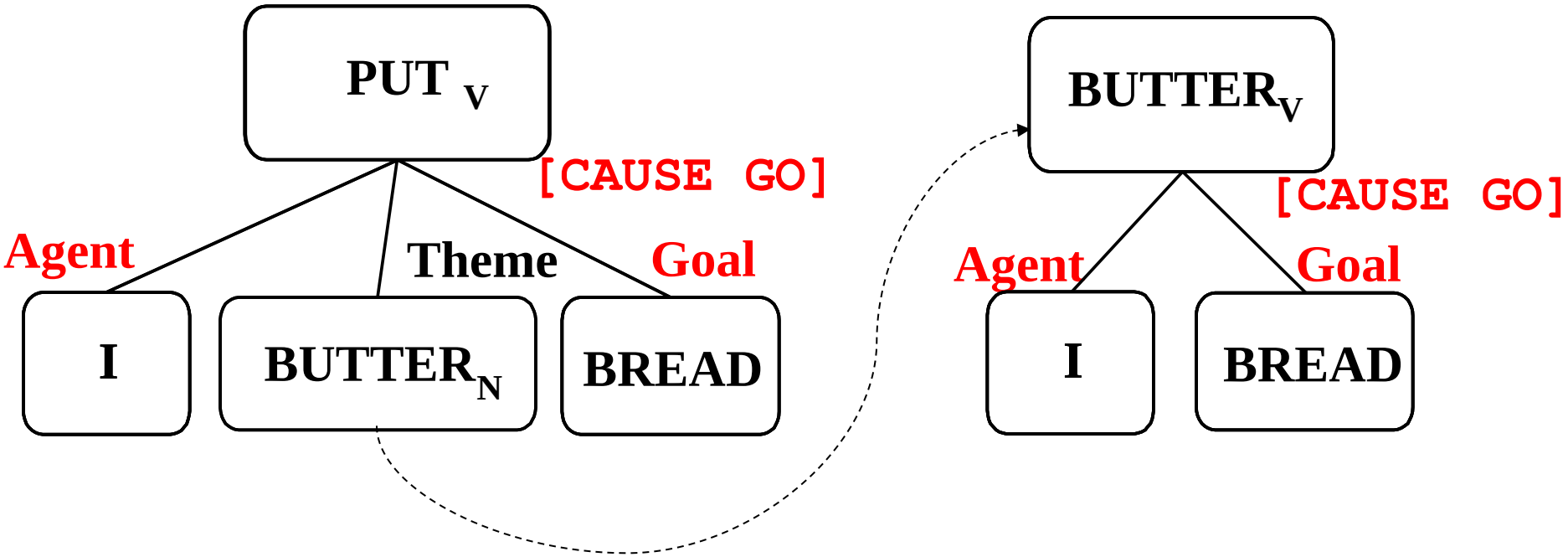
- Motivating Question: Can we inject statistical techniques into linguistically motivated MT?
- Using "approximate Interlingua" for MT
  - Tap into richness of deep target-language resources
    - Linguistic Verb Database (LVD) http://clipdemos.umiacs.umd.edu/englcslex/
    - CatVar database (CATVAR) http://clipdemos.umiacs.umd.edu/catvar/
- Constrained overgeneration
  - Generate multiple linguistically-motivated sentences
  - Statistically pare down results

[Work with Nizar Habash and Christof Monz, 2009]

# GHMT Example



# 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 → 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)

# What is paraphrase? (Madnani, Dorr, 2009)

Paraphrase = alternative surface form expressing the same semantic content as the original form, at one of three levels:

**Lexical**: Individual lexical items having same/similar meaning, i.e., synonyms such as *<correct, fix>.* Also, hypernyms: *<say, reply>*

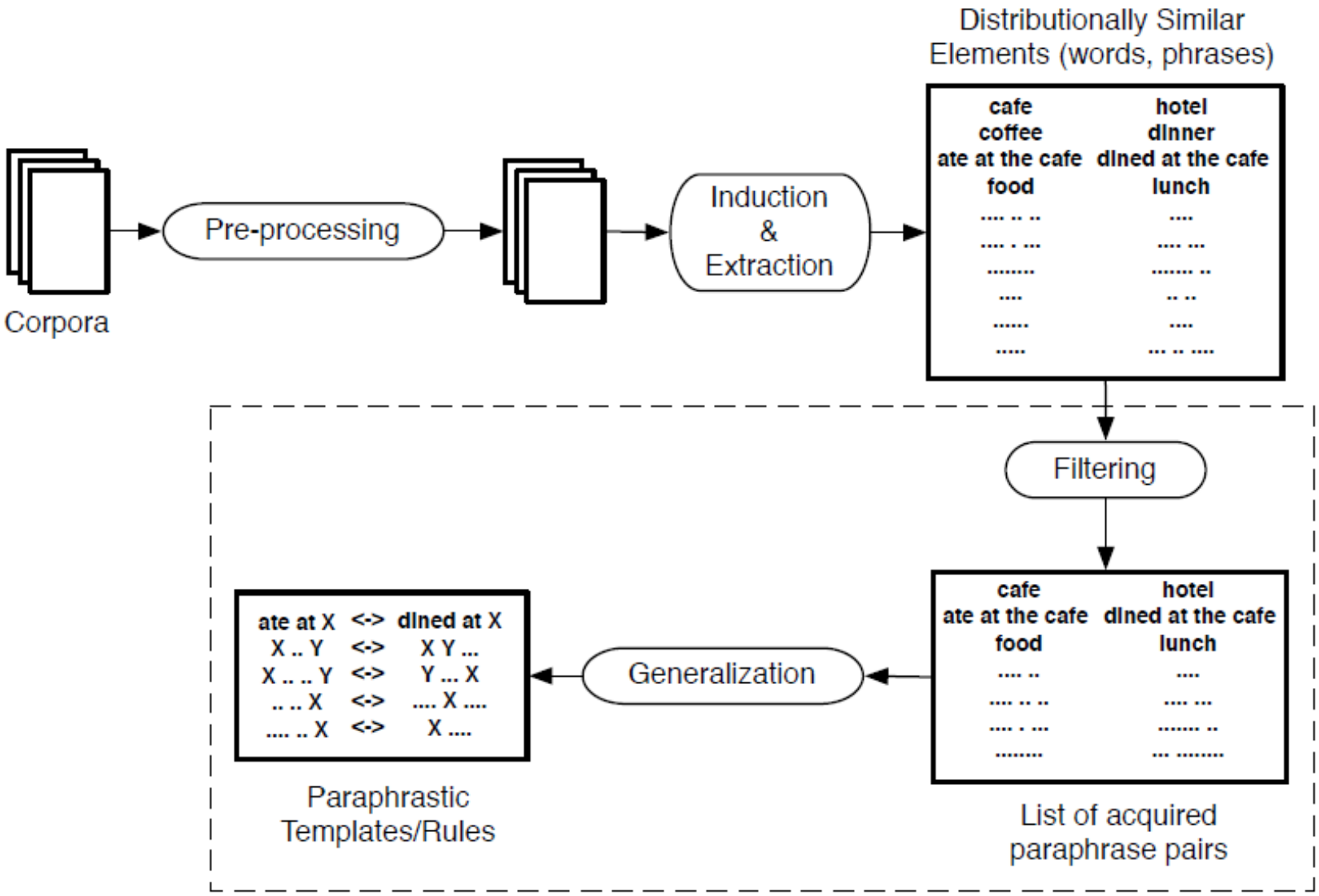
**Phrasal**: Phrasal fragments sharing the same semantic content, e.g., *<work on, soften up>.* Also, variable-ized forms: *<Y was built by X, X is the creator of Y>*

**Sentential**: Two sentences that represent the same semantic content, e.g., *<I finished my work, I completed my assignment>.* Also, more complicated forms: *<He needed to make a quick decision in that situation, The scenario required him to make a split-second judgment>*

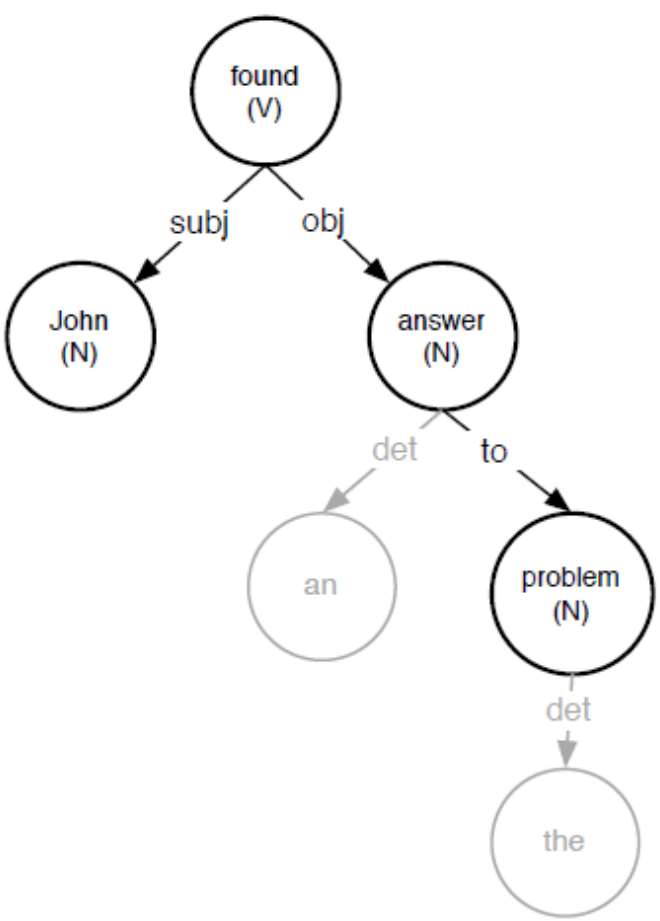

# Two MT-Related Applications for Paraphrase (Madnani, Dorr, 2009)

- Expanding Sparse Human Reference Data for MT Eval
  - N-gram overlap with human-generated reference (Papineni et al. 2002), but single reference translation cannot capture all possible verbalizations that convey same semantic content.
  - Penalization of non-overlapping outputs with same meaning, e.g., *<consider entire community, bear in mind community as a whole>*
  - Solution: (1) Multiple references – expensive! (2) Take into account paraphrases in reference translations (Zhou et al '06).
- Statistical MT Improvements
  - Use automatically induced paraphrases to improve statistical phrase-based MT system (Callison-Burch et al '09).
    - Divide sentence into phrases and translate each phrase from table look-up, inserting paraphrases for untranslatable source phrases.
  - Reference sparsity in MT parameter tuning (Madnani&Dorr '08, '13).
    - Expand single-reference tuning sets by including paraphrases.
    - More recently: Generate targeted paraphrases

# General Architecture for Paraphrasing from Distributional Similarity (Madnani, Dorr, 2009)



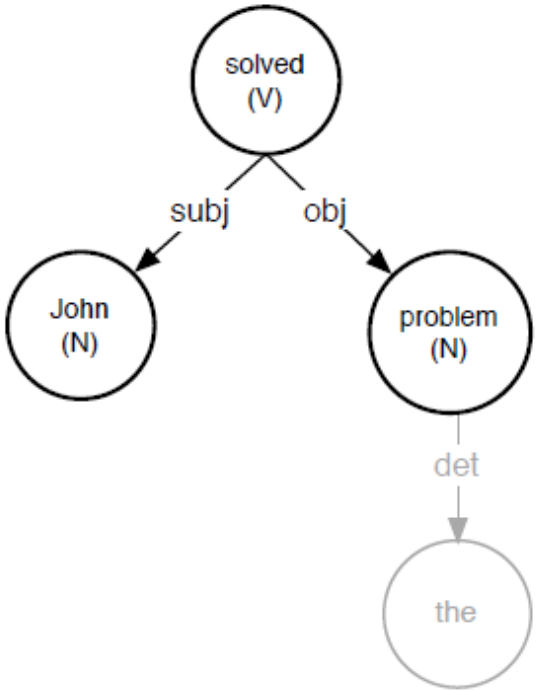
# Lin and Pantel (2001)




X Y

N:subj:V <- find-> V:obj:N -> answer -> N:to:N

"**X** found answer to **Y**"

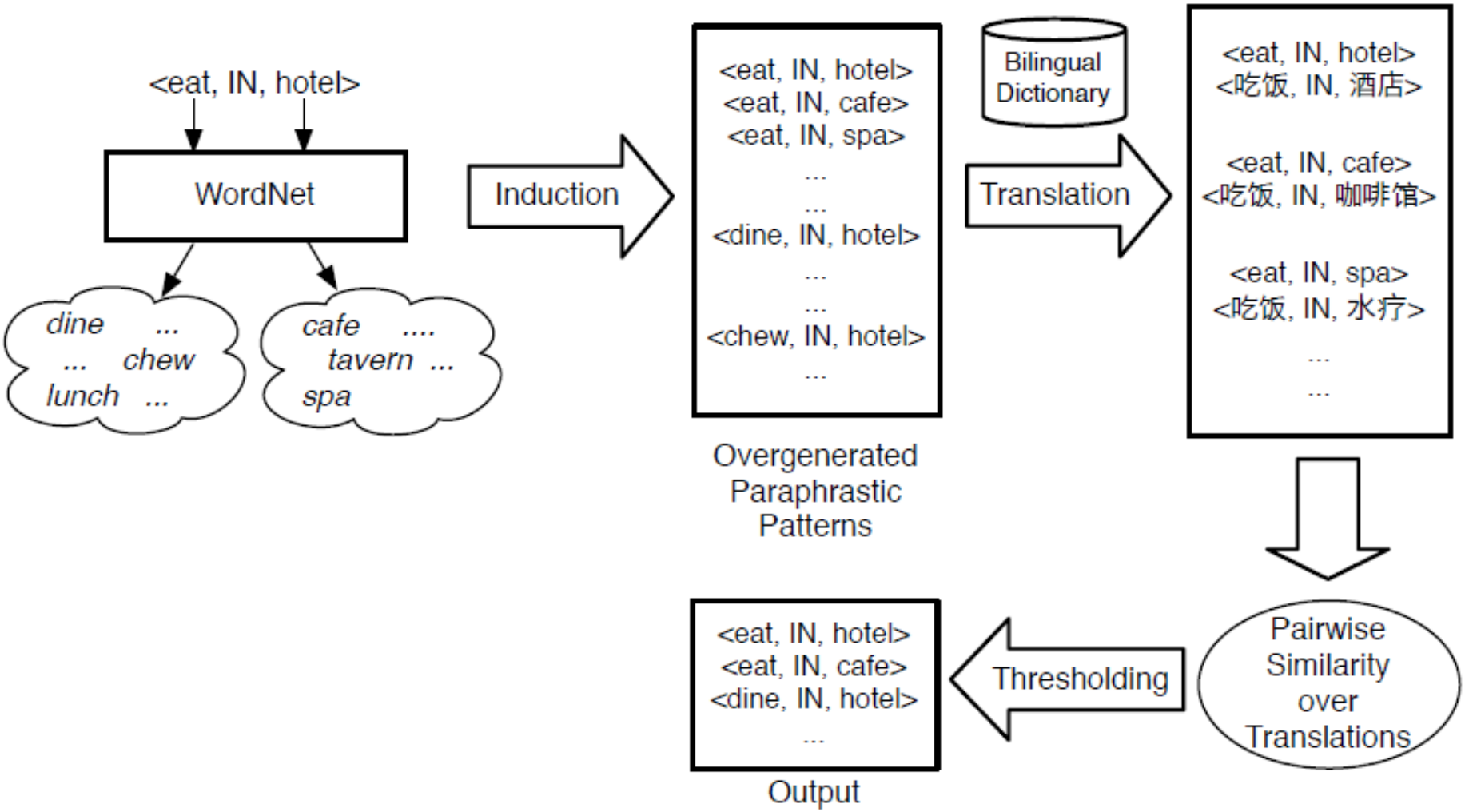



X Y

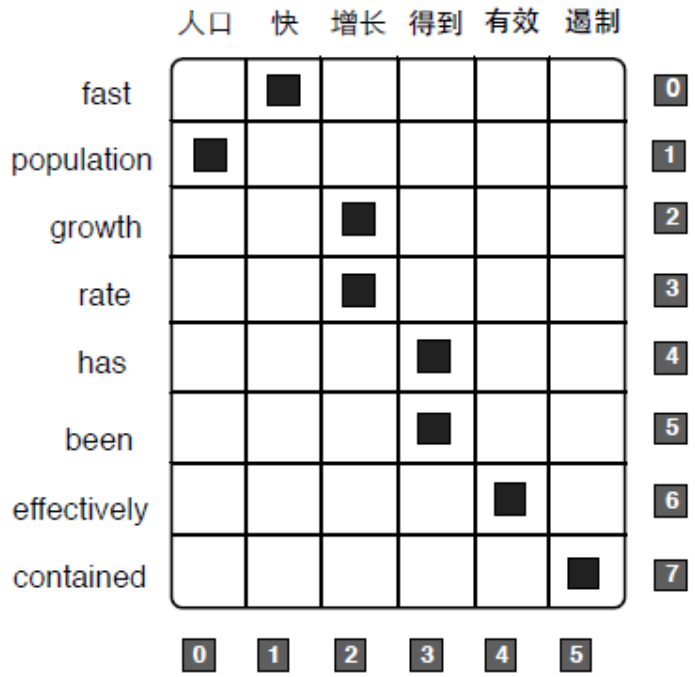
N:subj:V <- solve-> V:obj:N

"**X** solved **Y**"

# What about using a foreign language to compute distributional features? (Wu and Zhou, 2003)

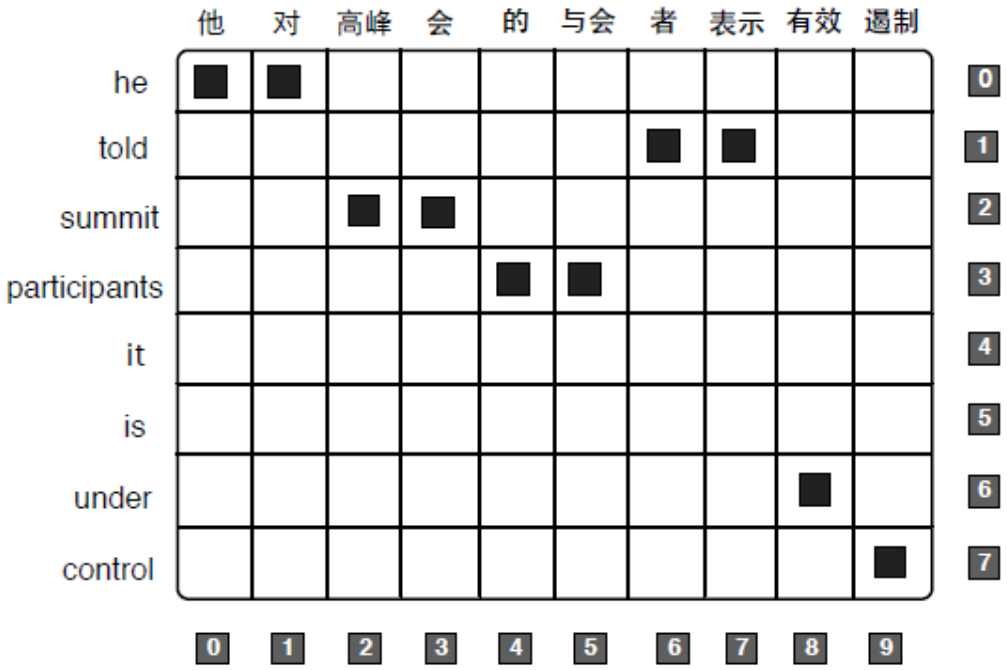


# Leveraging Bilingual Corpora for Inducing Paraphrases (Bannard & Callison-Burch, 2005)




**Extracted phrases**

```
(0,0) x (1,1) → <人口, population>
(1,1) x (0,0) → <快, fast>
(2,2) x (2,3) → <增长, growth rate>
...
...
(4,5) x (6,7) → <有效 遏制, effectively contained>
...
...
```




**Extracted phrases**

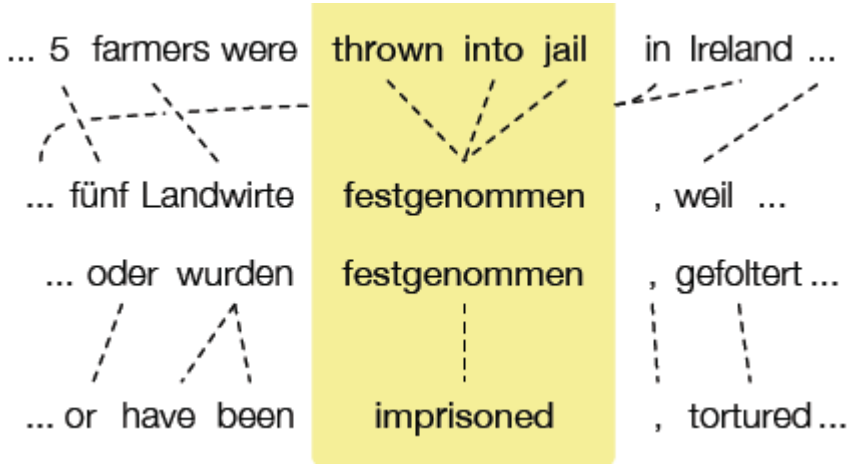
```
(0,1) x (0,0) → <他 对, he>
(6,7) x (1,1) → <者 表示, told>
(2,3) x (2,2) → <高峰 会, summit>
...
...
(8,9) x (6,7) → <有效 遏制, under control>
...
...
```

# Bannard and Callison-Burch (2005) vs. Wu and Zhou (2003), and Addressing "Noise"

- Both rely on secondary language to provide cues for paraphrase generation
  - Wu and Zhou: Reply on pre-compiled bilingual dictionary to discover cues
  - Bannard and Callison-Burch: An entirely data-driven discovery process using SMT alignment techniques.
- Madnani and Dorr (2009): Paraphrasing via bilingual corpora relies on word alignment that are often noisy. Using Arabic as pivot, found two categories of noise due to incorrect alignments:
  - Morphological variants. *<ten ton, ten tons>, <caused clouds, causing clouds>*.
  - Approximate Phrasal Paraphrases. Only shared partial semantic content. *<accounting firms, auditing firms>*
- Callison-Burch (on DEFT project) proposed an improvement that places an additional syntactic constraint on the phrasal paraphrases extracted via the pivot-based method from bilingual corpora
  - Using this constraint leads to a significant improvement in the quality of the extracted paraphrases.
  - Requires that the extracted paraphrase be of the same syntactic type as the original phrase.
  - Estimating the paraphrase probability now requires incorporation of syntactic type.

# A New Paraphrase Resource (NAACL 2013)

- PPDB: The Paraphrase Database (Ganitkevitch, Van Durme, Callison-Burch)
  - Collection of ranked English and Spanish paraphrases [DARPA's DEFT Program]
  - URL: paraphrase.org
- Built via two steps:
  - Extracting lexical, phrasal, and syntactic paraphrases from large bilingual parallel corpora (with associated paraphrase probabilities).




  - Computing distributional similarity scores for each of the paraphrases using the Google ngrams and the Annotated Gigaword corpus.
    - Uses n-gram features, position-aware and POS features, dependency link features, and other syntactic features

thrown into jail

| | | |
|---|---|---|
| arrested | be thrown in prison | arrest |
| detained | been thrown into jail | cases |
| imprisoned | being arrested | custody |
| incarcerated | in jail | maltreated |
| jailed | in prison | owners |
| locked up | put in prison for | protection |
| taken into custody | were thrown into jail | thrown |
| thrown into prison | who are held in detention | |

# What's missing? Inferring relationships, intentions, and entailment from informal communications

- Patterns of interaction reflect social situation (who has power, who has status) [Passonneau & Rambow, 2009]
  - Patterns of interaction (taken together with modality/confidence) reveal implicit relationships and underlying intentions
- 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].
- Paraphrase recognition for Textual Entailment and Similarity [Several papers in (NA)ACL 2013, *SEM 2013]

# 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)
- Semantically-Informed MT and Evaluation
  - Generation-Heavy Hybrid MT - more robust across genre
  - System combination produces best results
  - Automatic Paraphrasing for MT and Evaluation
- Inferring Relations/Intentions from informal communication
  - Requires deeper linguistic knowledge
  - Potential for hybrid linguistic/statistical approach to inference

# Questions?

bonnie.dorr@darpa.mil