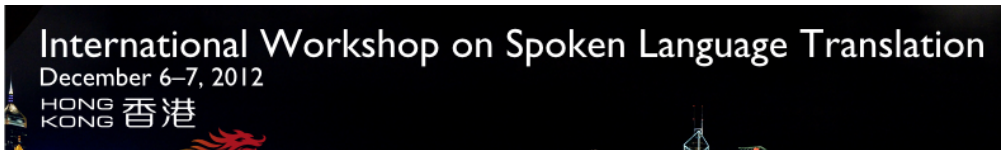
Active Error Detection and Resolution for Speech-to-Speech (S2S) Translation

Rohit Prasad
Rohit Kumar
Sankaranarayanan Ananthakrishnan
Wei Chen
Sanjika Hewavitharana
Matthew Roy
Frederick Choi
Aaron Challenner
Enoch Kan
Arvind Neelakantan
Prem Natarajan



Raytheon
BBN Technologies

Limitations of S2S Translation Systems

- **Serial integration of automatic speech recognition (ASR), Machine Translation (MT) & Text-to-Speech (TTS)**

- **Each component generates and propagates various types of errors**
  - **ASR issues (OOV words, homophones, mispronunciations)**
  - **Translation errors due to word sense ambiguities and idioms**
  - **Miscellaneous problems (e.g. fragments due to user error)**

- **Systems lack the ability to detect and recover from critical errors that impede communication flow**
  - **Error detection and recovery is largely the users' prerogative**

Research Goals

- **Improve S2S Translation Systems**
  - **Active Error Detection**
    - **Focusing on seven error types** (Stallard et. al.,2008; DARPA BOLT)

Problem Type	Example
Out-of-Vocabulary (OOV) Names	User: My name is Sergeant Gonzales. ASR: my name is sergeant guns all us
Out-of-Vocabulary (OOV) Words	User: The utility prices are extortionate. ASR: the utility prices are extort unit
Word Sense Ambiguities	User: Does the town have enough tanks. Ambiguous Senses: armored vehicle | storage unit
Homophones	User: Many souls are in need of repair. Ambiguous Homophones: soles | souls
Mispronunciation	User: Have people been harmed by the water when they wash. ASR: Have people been harmed by the water when they worse
Incomplete Utterances	ASR: Can you tell me what these
Idiomatic Phrases	User: We will go the whole nine yards to help. Idiom: the whole nine yards

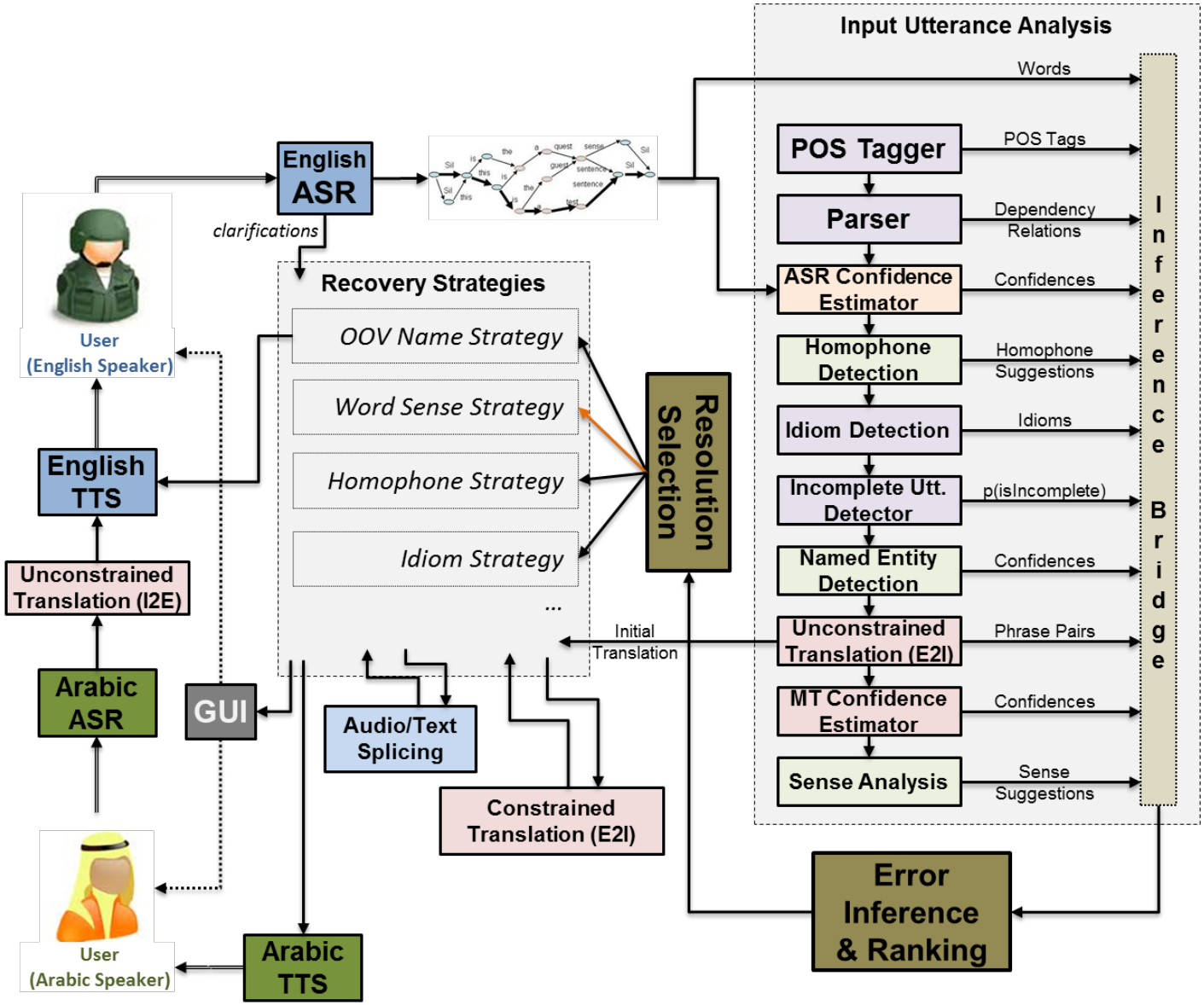
  - **Interactive Error Resolution**
    - **Transform systems from *passive conduits* of information transfer to *active participants***

Approach

- **Active Error Detection**
  - Errors are *detected* through a series of analysis
    - Analysis of both input utterance and translation output
    - Interaction context not used (currently)
  - Errors are *localized* to provide relevant feedback to user
  - Errors are *prioritized* to focus resolution on most severe errors

- **Interactive Error Resolution**
  - Mixed-Initiative Error Resolution
    - Attempt automatic error recovery
    - Engage the users: Only using English language speaker (currently)
  - Robust & Efficient Error Resolution Strategies
    - Users may override system in case of false alarms
    - *(Expert)* Users can still voluntarily identify & correct errors

Approach: System Architecture



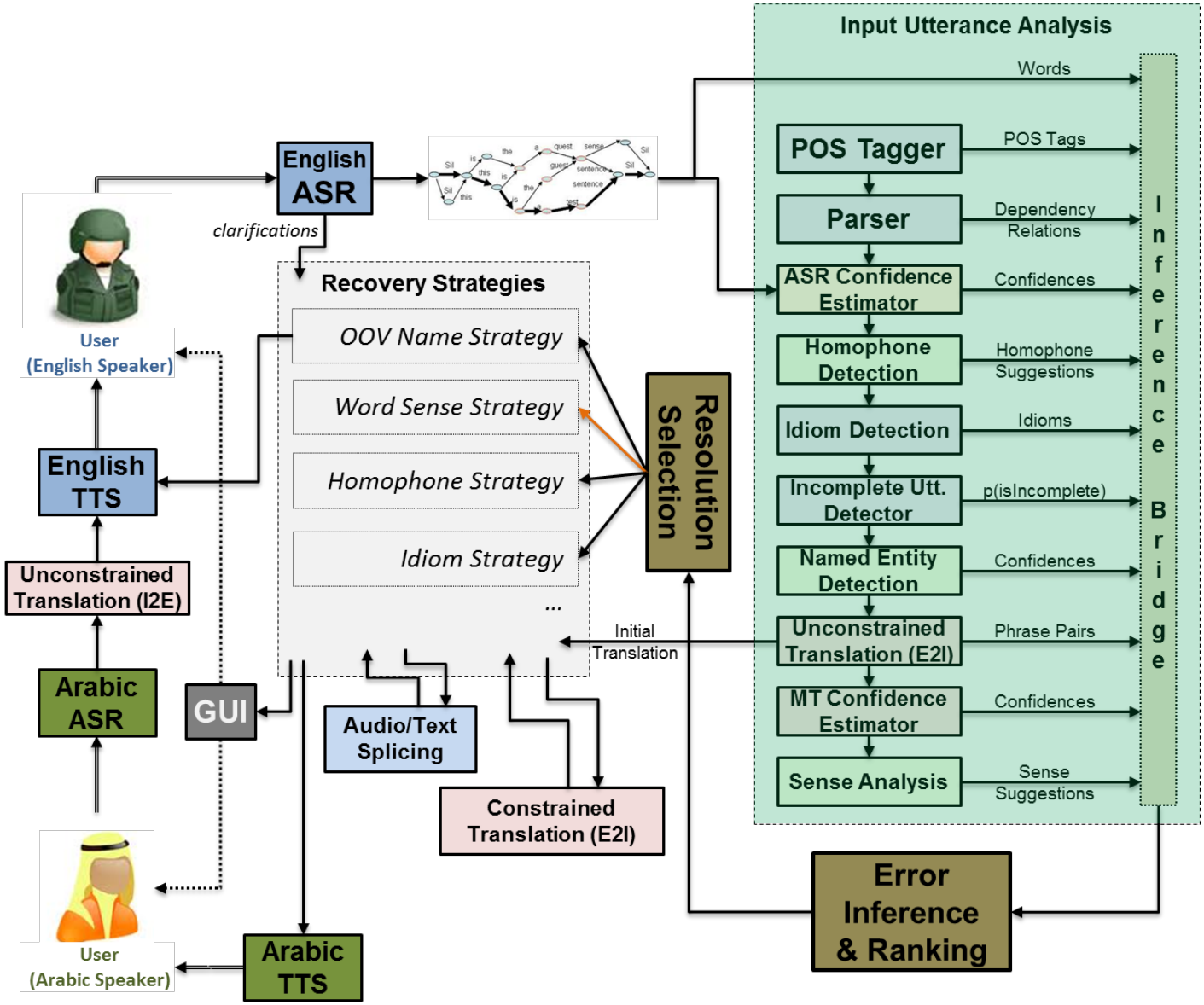
Core Components

- **Automatic Speech Recognition (ASR)**
  - BBN Byblos ASR
  - English AM: Trained on DARPA TRANSTAC corpus (150 hours)
  - English LM: Trained on 5.8m utterances/60m words (Vocab: 38k)
  - WER: 11%

- **Statistical Machine Translation (SMT)**
  - DARPA TRANSTAC English-Iraqi parallel corpus
    - 773k sentence pairs, 7.3m words
  - E2I BLEU: 16.1

- **Text-to-Speech (TTS)**
  - SVOX TTS Engine

Approach: System Architecture

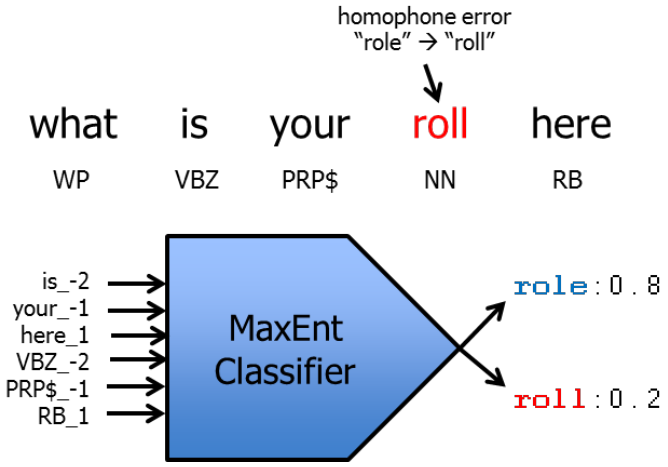


OOV Named-Entity Detection

- Gonzales ➔ recognized as ➔ guns all us
- MaxEnt classifier: Named-Entity Recognition (NER)
  - 250k utterances, 4.8m words, 450k names
- Rich Contextual Features
  - Lexical features (n-grams)
  - Syntactic features (part of speech)
  - Trigger words
- Fusing NER posteriors and ASR confidence scores
  - Early and late fusion techniques explored
- Detection Rate (Recall):
  - In-Domain Utterances: 40.5%
    - Additional 19.9% of OOV NEs detected by Error Span detector

Homophone Error Correction

- **Targeted Error Correction**
  - **MaxEnt classifier with context and dependency features to predict & correct homophone variants**
  - **Strong, locally discriminative LM**
- **Offline Evaluation**
  - **95.7% correction rate on a corpus with single word substitution error**
  - **1.3% false corrections on a corpus with no homophone errors**



Word Sense Errors: 2-pronged approach

- **Predict sense labels for ambiguous English words**
  - **Pre-defined inventory of ambiguity classes and senses**
  - **Approach and features follow homophone corrector**
- **Offline evaluation on 110 ambiguity classes**
  - **73.7% majority sense prediction baseline accuracy**
  - **88.1% sense prediction accuracy with MaxEnt**

Sample confusion matrices for two ambiguity classes in the evaluation set

	additional	remote
additional	11	1
remote	1	12

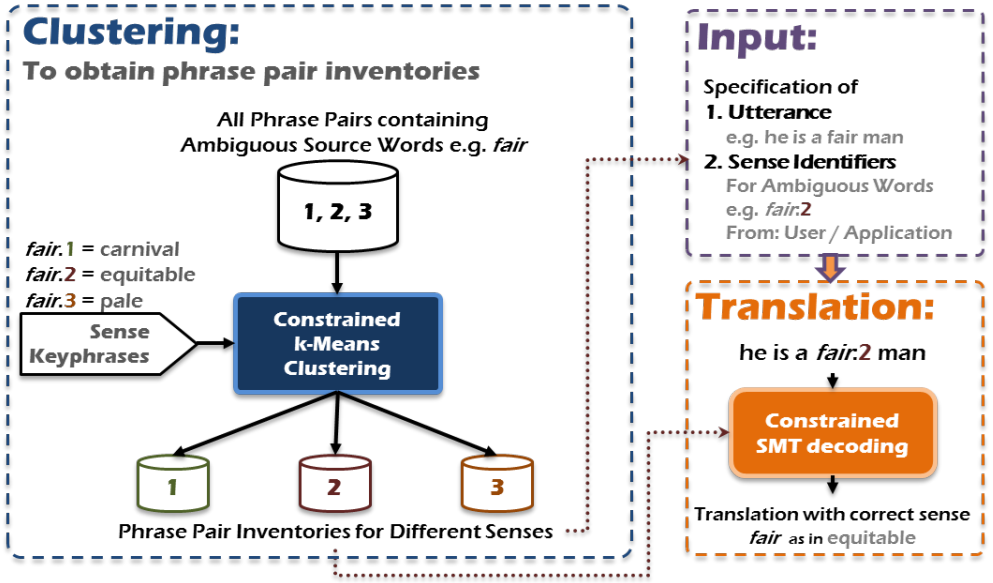
FURTHER = {further}

	record	currency
record	7	2
currency	0	5

NOTE = {note, notes}

Sense-Constrained SMT Decoding

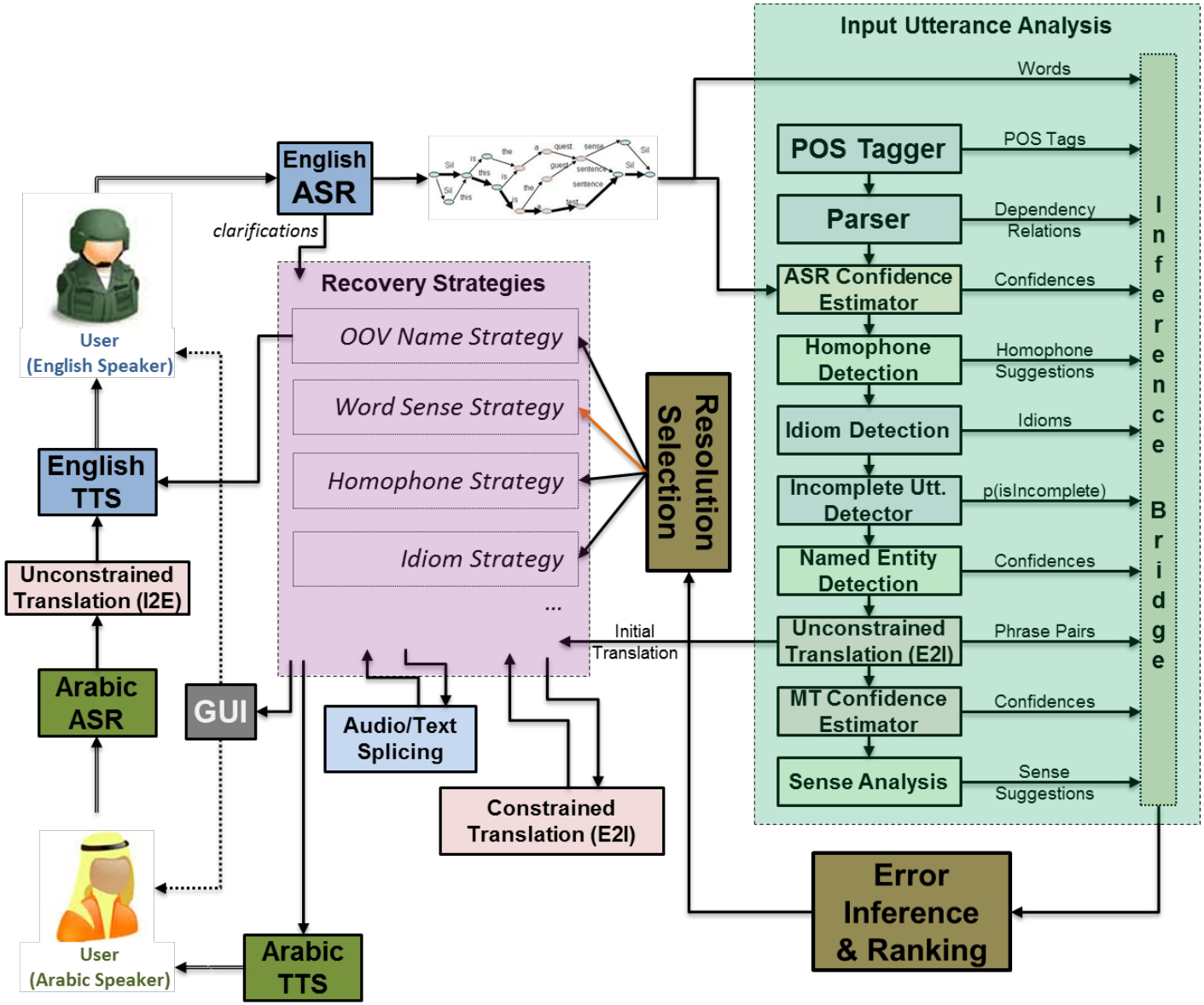
- Sense prediction does not guarantee correct translation
- Constrained SMT Decoding (dynamic pruning)
  - Apply phrase pairs from sense-specific partitions
  - Sense identifiers from MaxEnt predictor or user
- Generating phrase pair partitions
  - Novel semi-supervised approach
  - Constrained *k*-means clustering
  - Sense key-phrases used to seed constraints



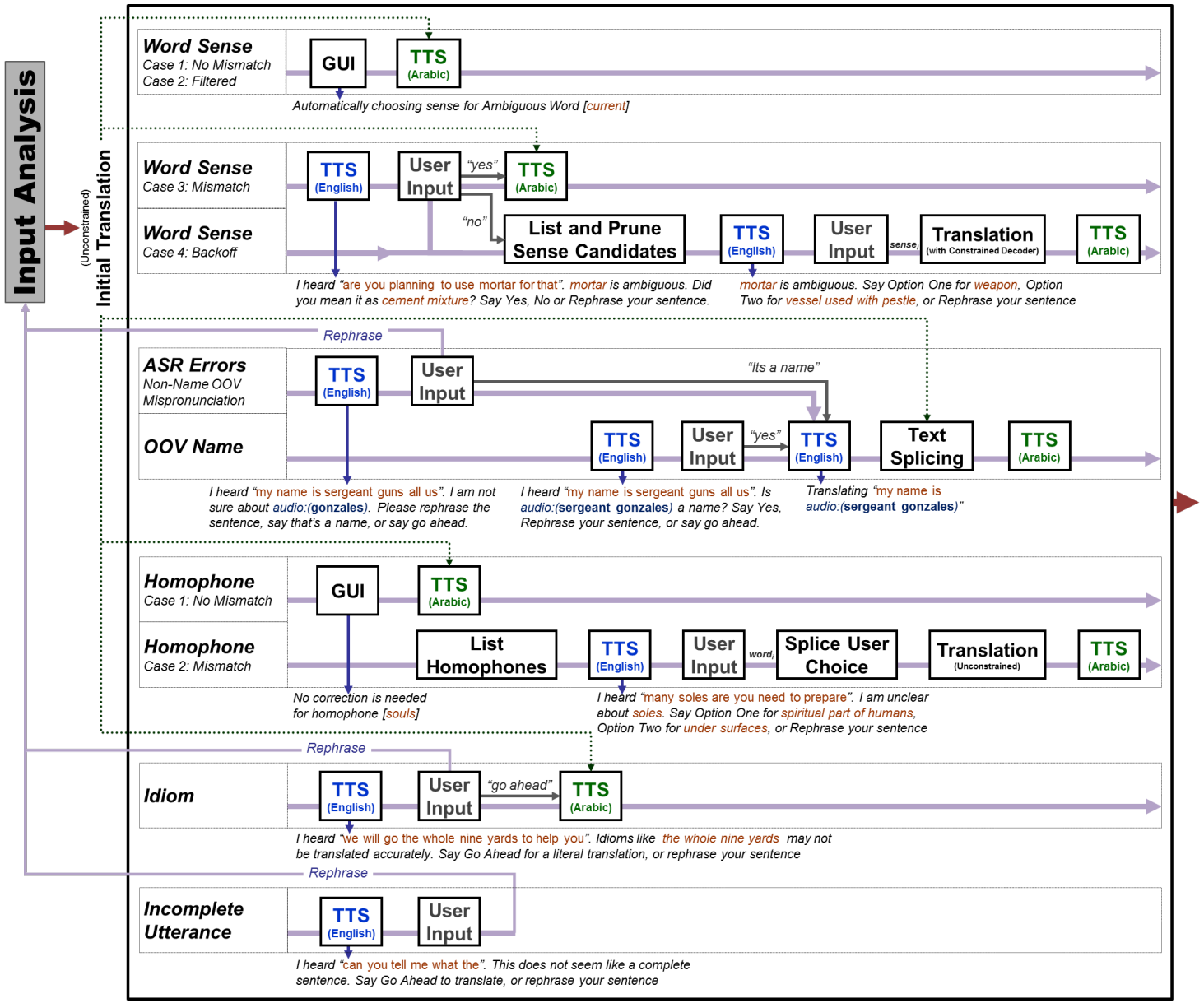
Other Detectors: Idioms, Fragments, Error Spans

- **Idiom Detection**
  - MaxEnt classifier trained on 20,000 idioms
  - Precision = 71.7%, Recall = 22.4%
- **Incomplete Utterance Detection**
  - Utterance-level MaxEnt classifier trained on unsupervised, automated fragment simulator
  - Precision = 82.5%, Recall = 41.9%
- **Error Span Detector**
  - Combines ASR & MT Confidence
  - Designed to catch words that will result in poor translation
  - Helps with detection of Unseen Translation phrases, User mispronunciations, OOVs & Other ASR errors

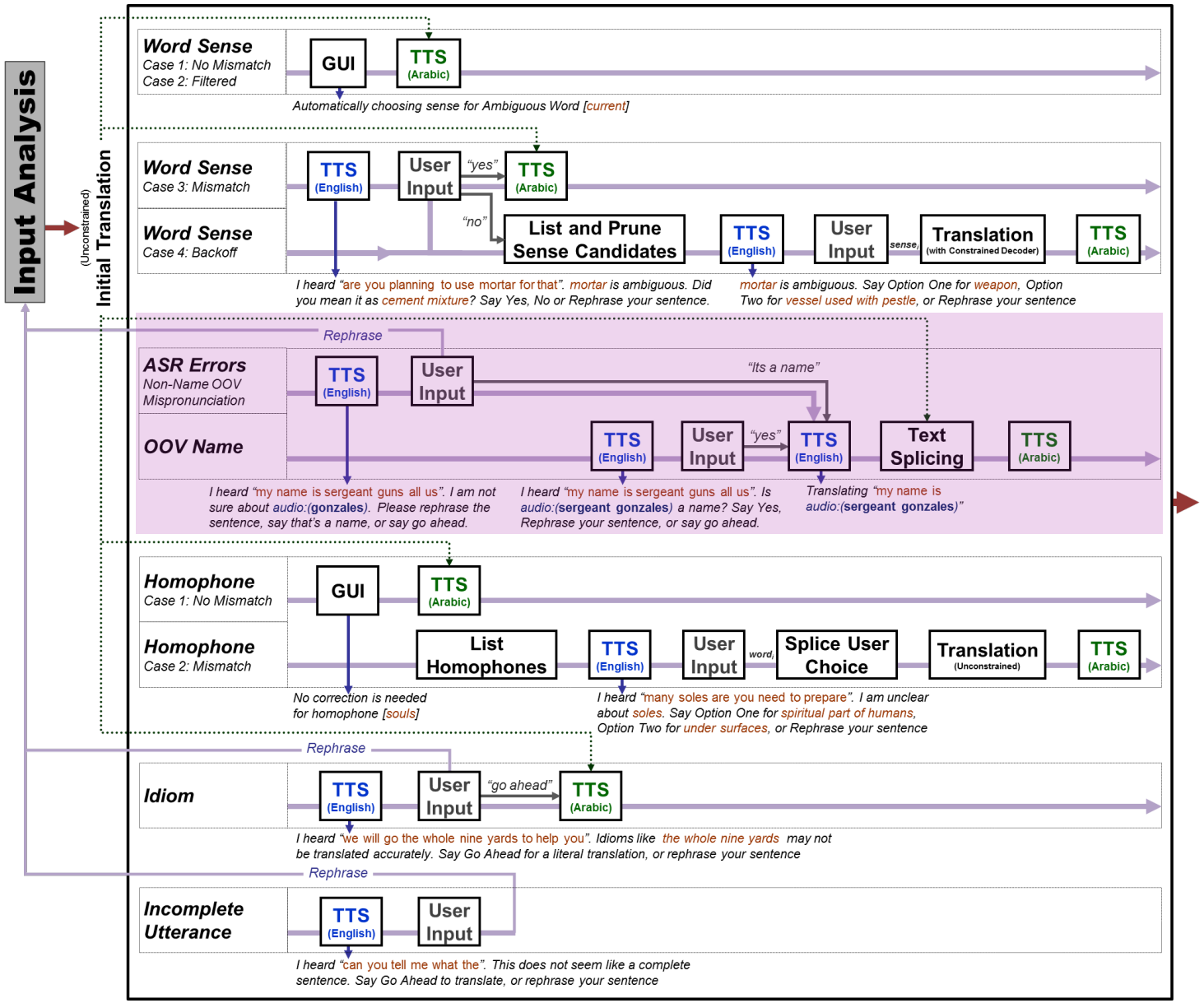
Approach: System Architecture



Error Resolution Strategies: Summarized

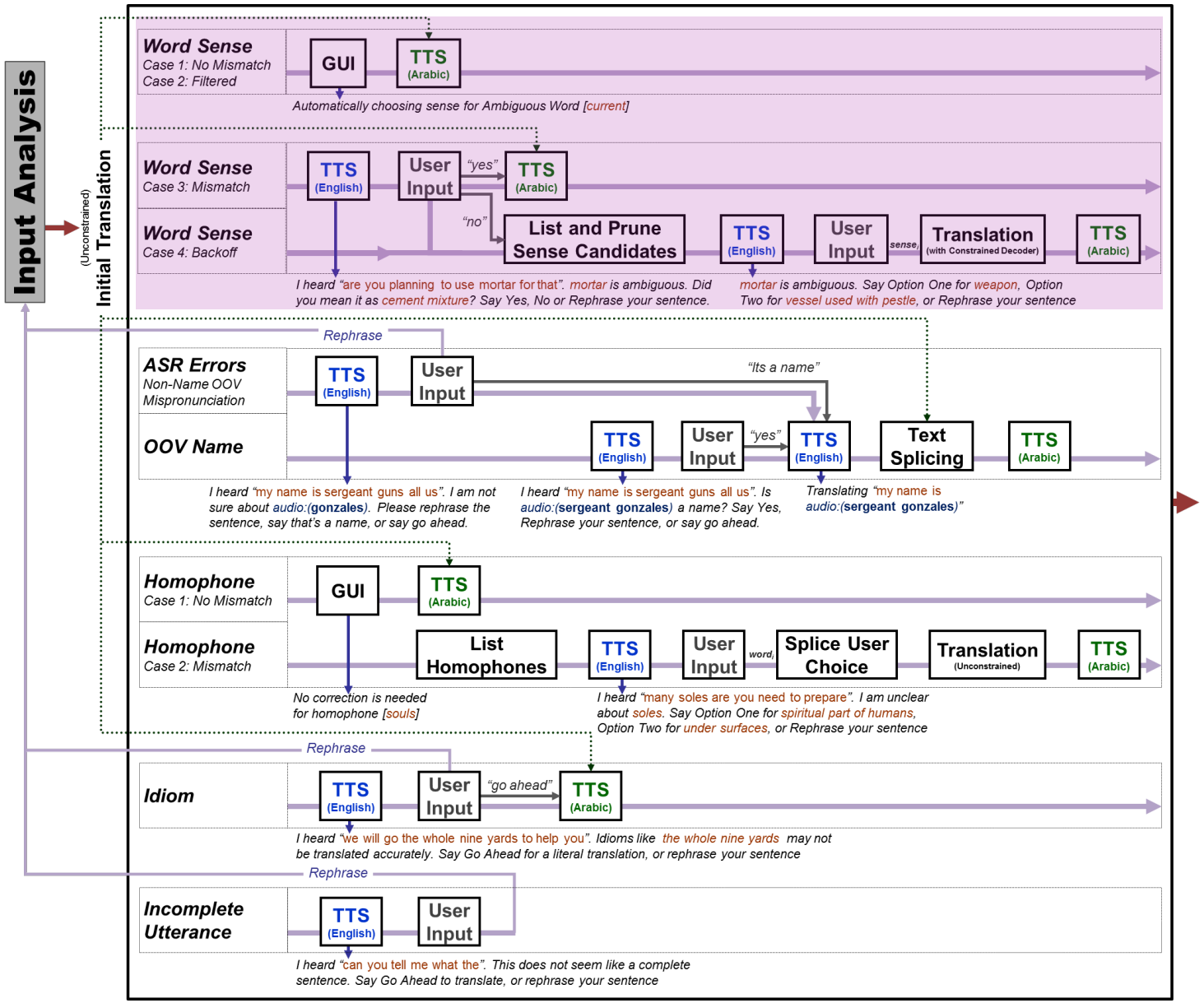


Error Resolution Strategies: Summarized

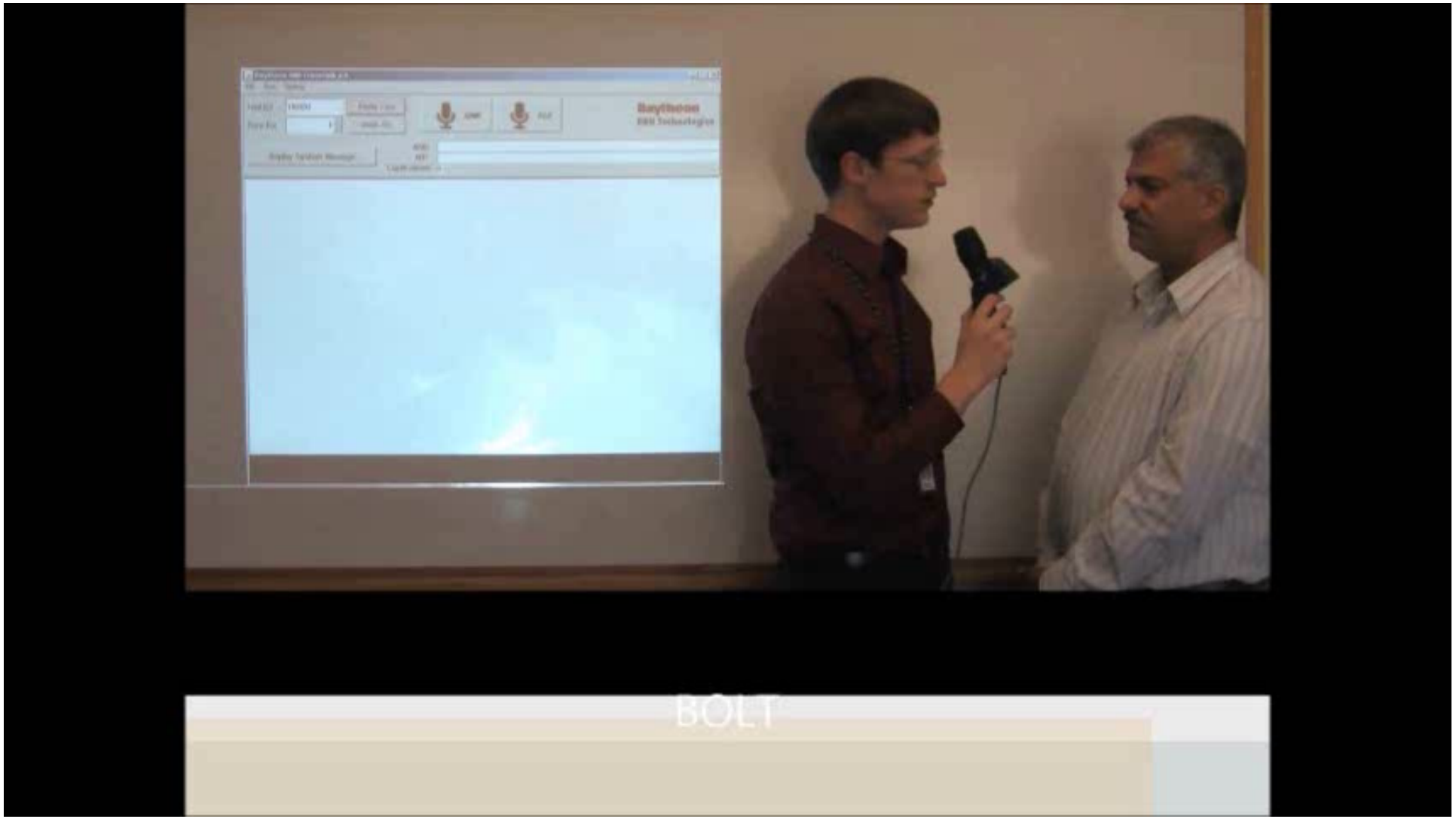


OOV Named Entity Error Resolution: Example

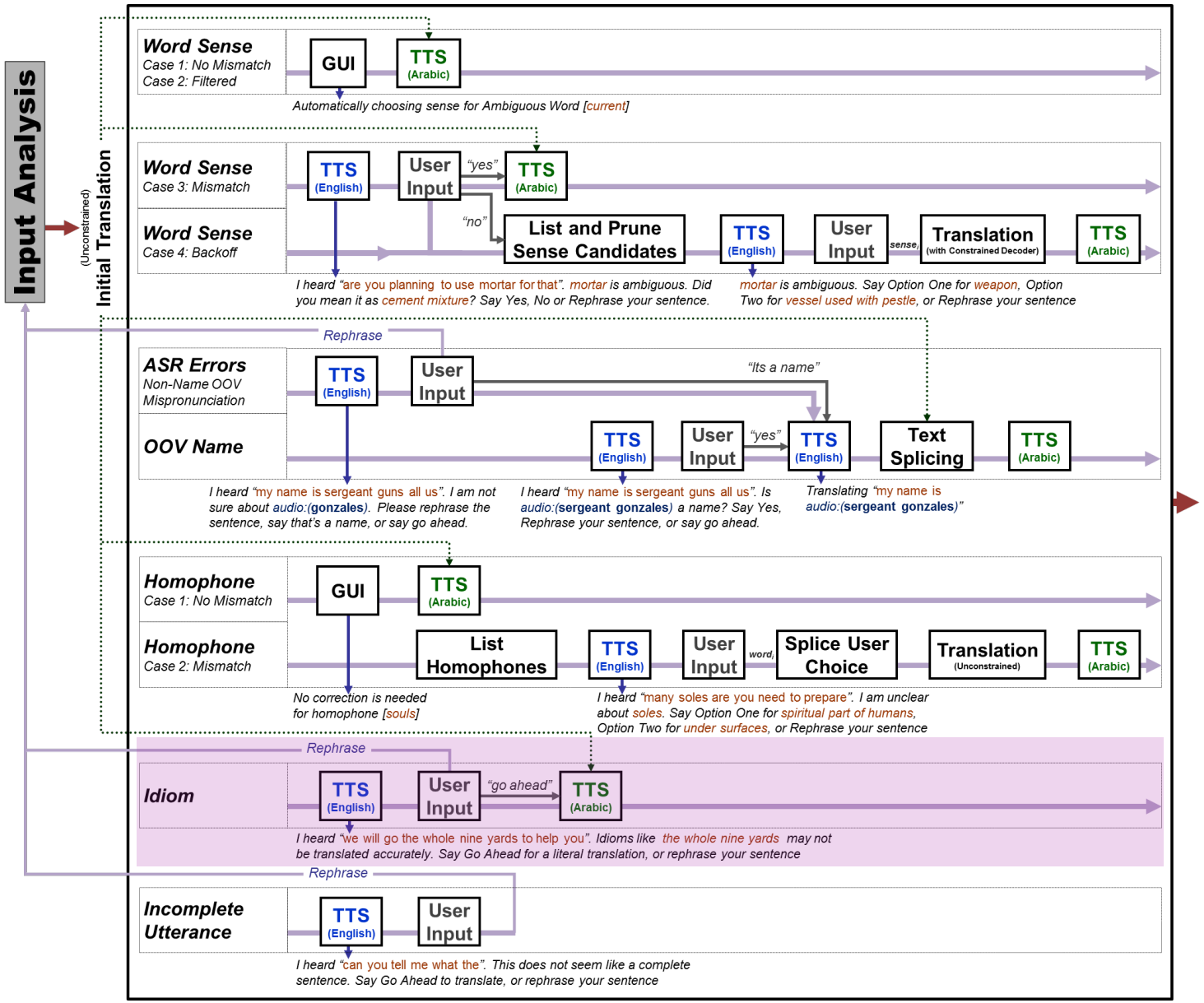
Error Resolution Strategies: Summarized



Word Sense Error Resolution: Example



Error Resolution Strategies: Summarized



Idiom Error Resolution: Example

Preliminary Evaluation: Methodology

- **20 scenarios**
  - **Consists of 5 starting utterances**
    - **Designed to elicit errors**
    - **Example Scenario:**

      Sir, I need to quiz you about your comings and goings

      Do you own the dealership in Hebeb

      We've heard of insurgent fliers being seen around here

      Do your competitors have suspicious contacts

      It sounds like there is a kernel of truth to your story

  - **Speaker speaks 1 utterance**
    - **Engages in clarification with system**

- **Speakers trained to use the system for 5 scenarios**

Preliminary Evaluation: Results

Intended Error	%Correct	%Recoverable
OOV-Name	41.7	75.0
OOV-Word	37.8	75.6
Word Sense*	16.7	16.7
Homophone*	31.3	50.0
Mispronunciation	60.0	60.0
Idiom	0.0	0.0
Incomplete	20.0	80.0
All	33.0	59.2

Error Detection Accuracy

- %Correct = %utterances where detected errors is the same as intended error
- %Recoverable = %utterances where detected error allows recovery from intended error

High Level Concept Transfer for Erroneous Concept

- Initial Transfer (before clarification)
- Final Transfer (after clarification)
- Recovery = (Final Transfer – Initial Transfer)

Intended Error	Initial Transfer	Final Transfer	Change
OOV-Name	8.3	41.7	33.4
OOV-Word	6.5	43.5	37.0
Word Sense	22.2	55.6	33.4
Homophone	26.7	33.3	6.6
Mispronunciation	20.0	40.0	20.0
Idiom	0.0	50.0	50.0
Incomplete	0.0	100.0	100.0
All	12.6	46.6	34.0

Conclusions

- **Active Error Detection & Interactive Resolution shown to improve transfer of erroneous concepts by 34%**
  - **Baseline: 12.6% (worse for certain types of errors)**
    - **Necessary for S2S systems to implement such capabilities for robustness**
  - **Improved System only able to transfer 46.6% concepts**
    - **Large scope/need for improvement**
  - **Towards High Precision S2S Systems**
    - **Trade-off between improved concept transfer and user effort**
    - **Current Evaluation: 1.4 clarification turns on average**

- **Directions**
  - **2-way S2S Systems with Active Error Detection & Resolution**
    - **Engaging both the speakers in error recovery**
  - **Reducing false-alarms / Minimizing the cost of false-alarm**

SPARE SLIDES

Constrained SMT Decoding Evaluation

- **Offline evaluation of constrained decoding with sense-specific phrase pair inventories**
- **73 ambiguity classes with multiple senses in training data**
- **164 sentences covering all senses of each ambiguity class**
- **Hand-tagged sense labels for each instance**
- **Human evaluated translation of ambiguous word (yes/no)**

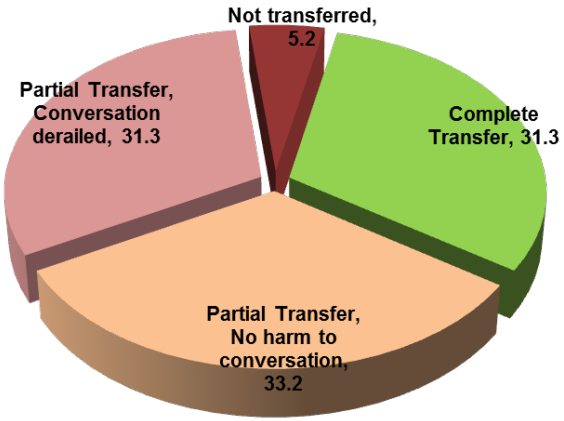
English input	Baseline translation	Constrained decoding
after our ***late*** leader died our town mourned for several weeks	bEd mAltnA ***mtJxr*** {*delayed*} AlqAQd mAt bldtnA km JsbwE	bEd mAltnA **AlmrHwm** {*deceased*} AlqAQd mAt bldtnA km JsbwE
this fifty pound ***note*** will cover the cost of dinner	hCA xmsyn ***mlAHZp*** {*remark*} rH ygTy tklfp AlERAG	hCA xmsyn **Alwrqp**{*bill*} rH ygTy tklfp AlERAG

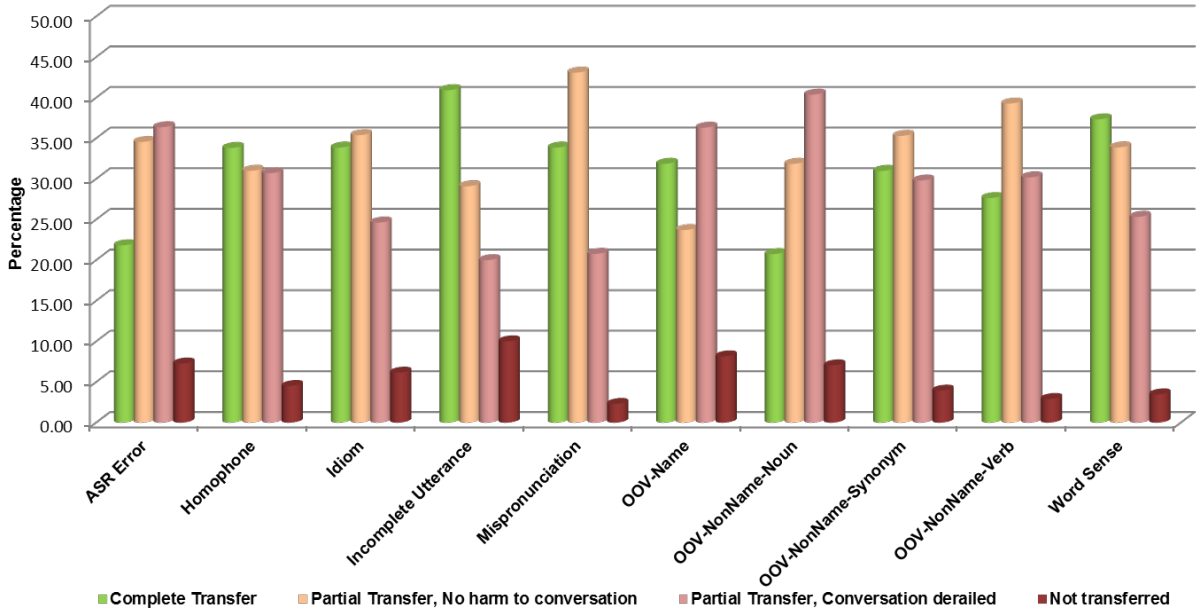
Examples illustrating translations of ambiguous words

	yes	no	unk
Baseline	95	68	1
Constrained	108	22	34
Improvement	13.7%	67.6%	n/a

Concept transfer accuracy for ambiguous words

BOLT Activity B/C Phase 1 Results





- 64% of the concepts (with targeted errors) are partially or completely transferred after clarification
  - Identifies and auto-corrects errors
  - System used only 1.3 clarification turns
- 62% of targeted errors are correctly identified by the system
- Transfer of erroneous concepts improved by 35% over the initial translation based on BBN's analysis of the demo logs

References

[BBN-IWSLT 2012] Rohit Prasad, Rohit Kumar, Shankar Ananthakrishnan, Wei Chen, Sanjika Hewavitharana, Matthew Roy, Frederick Choi, Aaron Challenner, Enoch Kan, Arvind Neelakantan, Prem Natarajan, **Active Error Detection and Resolution for Speech-to-Speech Translation**, Intl. Wksp. on Spoken Language Translation (IWSLT), 2012, Hong Kong

[BBN-Interspeech 2012] Rohit Kumar, Rohit Prasad, Shankar Ananthakrishnan, Aravind Vembu, Dave Stallard, Stavros Tsakalidis, Prem Natarajan, **Detecting OOV Named-Entities in Conversational Speech**, Interspeech 2012, Portland, Oregon

[ICASSP 2012] Shankar Ananthakrishnan, Stavros Tsakalidis, Rohit Prasad, Prem Natarajan, Aravind Vembu, **Automatic pronunciation prediction for text-to-speech synthesis of dialectal arabic in a speech-to-speech translation system**, ICASSP 2012, Kyoto, Japan

[BBN-NAACL 2013] Shankar Ananthakrishnan, Sanjika Hewavitharana, Rohit Kumar, Enoch Kan, Rohit Prasad, Prem Natarajan, **Semi-Supervised Word Sense Disambiguation for Mixed-Initiative Spoken Language Translation**, NAACL 2013 (*pending submission*)