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
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)

<table>
<thead>
<tr>
<th>Problem Type</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Out-of-Vocabulary (OOV) Names</td>
<td>User: My name is Sergeant Gonzales. ASR: my name is sergeant guns all us</td>
</tr>
<tr>
<td>Word Sense Ambiguities</td>
<td>User: Does the town have enough tanks. Ambiguous Senses: armored vehicle</td>
</tr>
<tr>
<td>Homophones</td>
<td>User: Many souls are in need of repair. Ambiguous Homophones: soles</td>
</tr>
<tr>
<td>Mispronunciation</td>
<td>User: Have people been harmed by the water when they wash. ASR: Have people</td>
</tr>
<tr>
<td>Incomplete Utterances</td>
<td>ASR: Can you tell me what these</td>
</tr>
<tr>
<td>Idiomatic Phrases</td>
<td>User: We will go the whole nine yards to help. Idiom: the whole nine yards</td>
</tr>
</tbody>
</table>

– 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 $\rightarrow$ recognized as $\rightarrow$ 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

<table>
<thead>
<tr>
<th></th>
<th>additional</th>
<th>remote</th>
</tr>
</thead>
<tbody>
<tr>
<td>additional</td>
<td>11</td>
<td>1</td>
</tr>
<tr>
<td>remote</td>
<td>1</td>
<td>12</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>record</th>
<th>currency</th>
</tr>
</thead>
<tbody>
<tr>
<td>record</td>
<td>7</td>
<td>2</td>
</tr>
<tr>
<td>currency</td>
<td>0</td>
<td>5</td>
</tr>
</tbody>
</table>

FURTHER = \{further\}

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

**Word Sense**
- Case 1: No Mismatch
- Case 2: Filtered

- Automatically choosing sense for Ambiguous Word [current]

**Word Sense**
- Case 3: Mismatch

- User Input
- TTS [Arabic]

- GUI
- TTS [Arabic]

**ASR Errors**
- Non-Name OOV Mispronunciation

- OOV Name

- User Input
- TTS [English]

- TTS [English]

- Text Splicing
- TTS [Arabic]

**Homophone**
- Case 1: No Mismatch

- GUI
- TTS [Arabic]

- No correction is needed for homophone [spoil]

**Homophone**
- Case 2: Mismatch

- List Homophones
- TTS [English]

- User Input
- TTS [English]

- Splice User Choice
- TTS [Arabic]

**Idiom**

- TTS [English]

- User Input
- "go ahead"

- TTS [Arabic]

- "I heard "many sides are you need to prepare". I am unclear about sides. Say Option One for spiritual part of humans, Option Two for under surfaces, or Rephrase your sentence"

**Incomplete Utterance**

- TTS [English]

- User Input

- TTS [Arabic]

- "I heard "can you tell me what the". This does not seem like a complete sentence. Say Go Ahead to translate, or rephrase your sentence"
Error Resolution Strategies: Summarized
OOV Named Entity Error Resolution: Example
Error Resolution Strategies: Summarized

- **Word Sense**
  - Case 1: No Mismatch
  - Case 2: Filtered

- **Initial Translation**
  - Automatically choosing sense for Ambiguous Word [current]

- **Word Sense**
  - Case 3: Mismatch
    - TTS (Arabic)
    - User Input
    - "yes"

- **List and Prune Sense Candidates**
  - TTS (Arabic)
  - User Input

- **Translation**
  - With Constained Decoder
  - TTS (Arabic)

- **ASR Errors**
  - Non-Name OOV Mispronunciation
  - OOV Name
    - I heard "my name is sergeant guns all us. It's a name."
    - I am not sure about audio (gonzales). Please rephrase the sentence, say that's a name, or say go ahead.

- **Homophone**
  - Case 1: No Mismatch
  - GUI
  - TTS (Arabic)

- **Homophone**
  - Case 2: Mismatch
    - No correction is needed for homophone [suit]
    - I heard "many soles are you need to prepare". I am unclear about soles. Say Option One for spiritual part of humans, Option Two for under surfaces, or Rephrase your sentence.

- **Idiom**
  - TTS (English)
  - User Input: "go ahead"

- **Incomplete Utterance**
  - TTS (English)
  - User Input: "can you tell me what the"
Word Sense Error Resolution: Example
Error Resolution Strategies: Summarized

- **Word Sense**
  - Case 1: No Mismatch
  - Case 2: Filtered

- **ASR Errors**
  - Non-Name OOV Mispronunciation

- **OOV Name**
  - I heard “my name is sergeant guns all us”. I am not sure about audio (gonzales). Please rephrase the sentence, say that’s a name, or say go ahead.

- **Homophone**
  - Case 1: No Mismatch
  - Case 2: Mismatch

- **Idiom**
  - I heard “we will go the whole nine yards to help you”. Idioms like the whole nine yards may not be translated accurately. Say Go Ahead for a literal translation, or rephrase your sentence.

- **Incomplete Utterance**
  - I heard “can you tell me what the”. This does not seem like a complete sentence. Say Go Ahead to translate, or rephrase your sentence.
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

<table>
<thead>
<tr>
<th>Intended Error</th>
<th>%Correct</th>
<th>%Recoverable</th>
</tr>
</thead>
<tbody>
<tr>
<td>OOV-Name</td>
<td>41.7</td>
<td>75.0</td>
</tr>
<tr>
<td>OOV-Word</td>
<td>37.8</td>
<td>75.6</td>
</tr>
<tr>
<td>Word Sense*</td>
<td>16.7</td>
<td>16.7</td>
</tr>
<tr>
<td>Homophone*</td>
<td>31.3</td>
<td>50.0</td>
</tr>
<tr>
<td>Mispronunciation</td>
<td>60.0</td>
<td>60.0</td>
</tr>
<tr>
<td>Idiom</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>Incomplete</td>
<td>20.0</td>
<td>80.0</td>
</tr>
<tr>
<td>All</td>
<td>33.0</td>
<td>59.2</td>
</tr>
</tbody>
</table>

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)

<table>
<thead>
<tr>
<th>Intended Error</th>
<th>Initial Transfer</th>
<th>Final Transfer</th>
<th>Change</th>
</tr>
</thead>
<tbody>
<tr>
<td>OOV-Name</td>
<td>8.3</td>
<td>41.7</td>
<td>33.4</td>
</tr>
<tr>
<td>OOV-Word</td>
<td>6.5</td>
<td>43.5</td>
<td>37.0</td>
</tr>
<tr>
<td>Word Sense</td>
<td>22.2</td>
<td>55.6</td>
<td>33.4</td>
</tr>
<tr>
<td>Homophone</td>
<td>26.7</td>
<td>33.3</td>
<td>6.6</td>
</tr>
<tr>
<td>Mispronunciation</td>
<td>20.0</td>
<td>40.0</td>
<td>20.0</td>
</tr>
<tr>
<td>Idiom</td>
<td>0.0</td>
<td>50.0</td>
<td>50.0</td>
</tr>
<tr>
<td>Incomplete</td>
<td>0.0</td>
<td>100.0</td>
<td>100.0</td>
</tr>
<tr>
<td>All</td>
<td>12.6</td>
<td>46.6</td>
<td>34.0</td>
</tr>
</tbody>
</table>
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)

<table>
<thead>
<tr>
<th>English input</th>
<th>Baseline translation</th>
<th>Constrained decoding</th>
</tr>
</thead>
<tbody>
<tr>
<td>after our <em>late</em> leader died our town mourned for several weeks</td>
<td>bEd mAltnA <em>mtJxr</em></td>
<td>bEd mAltnA <em>AlmrHwm</em></td>
</tr>
<tr>
<td></td>
<td>{delayed} AlqAQd mAt</td>
<td>{deceased} AlqAQd</td>
</tr>
<tr>
<td></td>
<td>bldtnA km JsbwE</td>
<td>mAt bldtnA km JsbwE</td>
</tr>
<tr>
<td>this fifty pound <em>note</em> will cover the cost of dinner</td>
<td>hCA xmsyn <em>mlAHZp</em></td>
<td>hCA xmsyn <em>Alwrqp</em></td>
</tr>
<tr>
<td></td>
<td>{remark} rH ygTy tklfp AIERAG</td>
<td>{bill} rH ygTy tklfp AIERAG</td>
</tr>
</tbody>
</table>

Examples illustrating translations of ambiguous words

<table>
<thead>
<tr>
<th></th>
<th>yes</th>
<th>no</th>
<th>unk</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>95</td>
<td>68</td>
<td>1</td>
</tr>
<tr>
<td>Constrained</td>
<td>108</td>
<td>22</td>
<td>34</td>
</tr>
<tr>
<td>Improvement</td>
<td>13.7%</td>
<td>67.6%</td>
<td>n/a</td>
</tr>
</tbody>
</table>

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


