Speech Processing 11-492/18-492

Speech Translation
Speech Translation

◆ Three part systems
  ● ASR -> Translation -> TTS

◆ System configurations
  ● One way – phrasal
  ● One way – broadcast/lecture
  ● 1.5 way – phrasal with limited answers
  ● Two way – full two way
Machine Translation Technologies

- Phrasal
  - Phrase to phrase look up

- Template:
  - Template fillers, fixed translation

- Interlingua
  - Translation into meaning representation

- Statistical Machine Translation
  - From large collect of parallel text

- Classification base translation
  - Identify classes and deal directly with them
Choices in Translation

◆ Choose any two ...
  - High accuracy
  - Large vocabulary
  - Fully automatic

◆ Speech vs Text
  - Speech less clear than text
  - Less speech to train from
  - Needs to be real-time (probably)
Simple Translation

◆ Phrase to Phrase
  ● Greetings
  ● Do you need medical attention?
  ● Relatively easy to build, but limited use

◆ Template translations
  ● The next train leaves at TIME from gate GATE form PLACE
  ● Limited but still useful
**Interlingua**

- Translate sentences into standard form
- Generate sentences from standard form

**PROS:**
- Can do multiple languages easily
- Can be very accurate

**CONS**
- Designing universal interlingua is very hard
- Doesn’t do well when out of domain
Build probabilistic models from parallel text

Parallel text often available from

- Bilingual organizations
  - Governments, UN

- Relatively easy to collect
  - Requires translators rather than MT experts
## Learning from Parallel Text

<table>
<thead>
<tr>
<th>English</th>
<th>Lhiyohlili</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. The dog chases the cat</td>
<td>Ofi'at kowi'ã lhiyohli</td>
</tr>
<tr>
<td>2. The cat chases the dog</td>
<td>Kowi'at ofi'ã lhiyohli</td>
</tr>
<tr>
<td>3. The dog stinks</td>
<td>Ofi'at shoha</td>
</tr>
<tr>
<td>4. The woman loves the man</td>
<td>Ihooat hattakã hollo</td>
</tr>
<tr>
<td>5. I chase her/him</td>
<td>Lhiyohlili</td>
</tr>
<tr>
<td>6. She/he chases me</td>
<td>Salhiyohli</td>
</tr>
<tr>
<td>7. She/he dances</td>
<td>Hilha</td>
</tr>
</tbody>
</table>
# Learning from Parallel Text

| 1. Ofi'at kowi'ã lhiyohli | 1. The dog *chases* the cat |
| 2. Kowi'at ofi'ã lhiyohli | 2. The *cat chases* the dog |
| 3. Ofi'at shoha | 3. The *dog stinks* |
| 4. Ihooat hattakã hollo | 4. The *woman* loves the *man* |
| 5. Lhiyohlili | 5. I *chase* her/him |
| 6. Salhiyohli | 6. She/he *chases* me |
| 7. Hilha | 7. She/he *dances* |
Pros:
- Data collection doesn’t require MT experts
- Data driven
- Degrades gracefully when out of domain

Cons:
- Needs all language pairs
- Needs good/lots of data
- Hard to fix specific errors
Speech isn’t text
- Different style, hard to find lots of examples

Speech isn’t fluent
- False starts, hesitations, ungrammatical

ASR never makes errors 😊
One Way: Broadcast

- **One speaker**
  - Lecturer: can modify language model

- **Multiple speakers**
  - May be repeat speakers (News Anchor)
  - May had other noises: music etc
  - (TV programs)

- Doesn’t need to be real time (maybe)
Two Way: Dialog

- Users can detect own errors and correct
- Needs to be real time
- One user may be much more familiar
- How do you teach the other user
- Typically domain directed
Speech Technology Issues

- **ASR:**
  - Disfluencies, dialects, speaking style
  - Unfamiliarity with system

- **TTS:**
  - MT output isn’t always fluent
  - TTS says it anyway
  - Can be hard to understand
Speech Technology Issues

- **Spoken not Written Languages**
  - Arabic vs Arabic Dialects
  - Mixture of languages
  - Politeness levels
  - Gender in speech