Speech translation in human-to-human interaction: Skype Translator

InterAct25 – Baden-Baden

Chris.Wendt@microsoft.com

Group Program Manager – Machine Translation – Microsoft Research

July 14, 2016

@Tian500
Breaking down language barriers

Between humans

As natural as the air between us

photo.catwallpapers.info
Should the interpreter have a persona?
Redmond campus on Sept. 23, 2015. Photo by Brian Smale.
Why now?

Confluence of factors:

- Steady progress in MT quality over the last few years
  - Using vast amounts of data

Technological Leap in Speech Recognition

- Deep Learning (DNNs) – 33+% word error rate (WER) reduction over GMMs (Seide et al 2011)
  - From average of 30% down to 20%, in English
  - Now the improvement is above 42%
- More robust to noise, speaker variation, accents

Skype

- A global platform to put speech translation in the hands of 100s of Millions of users
Skype Translator: What is it?

1. Conversational Speech Recognition
2. TrueText
3. Machine Translation
4. Skype’s Worldwide network
5. A usable human interface

- Oh, I am also very good.
How people really speak

What person thought they said:

Yeah. I guess it was worth it.
→ Ja. Ich denke, es hat sich gelohnt.
→ 是的。我想这是值得的。

What they actually said:

Yeah, but um, but it was you know, it was, I guess, it was worth it.
→ Ja, aber ähm, aber es war, weißt du, es war, ich denke, es hat sich gelohnt.
→ 是的但是嗯，但你知道，它是，我猜，它是值得的。

Disfluency removal

More than just removing “um” and “ah”

TrueText is its own translation engine in the same language

Benutzten Sie dir vor?

Ich bitte Gurdeep, mir zu helfen.

Ähm Nein, ich meine ja, aber ich bin, habe ich es nie geschafft mich zuvor Benutzer äh ich werde Fragen gehen tief, um mir zu helfen


Benutztten Sie dir vor? Ich bitte Gurdeep, mir zu helfen.
There is a large perceived need for translation services in the US, across personal and business scenarios.

In the last 12 months, have you been in any situation where you wanted or needed to communicate with someone in another language that you are not currently able to speak proficiently or without help?

- Yes: 51%
- No

Reason for requiring translation:
- Both personal and business: 32%
- Personal reasons only: 51%
- Business reasons only: 17%

Source: S4: In the last 12 months, have you been in any situation where you wanted or needed to communicate with someone in another language that you are not currently able to speak proficiently or without help?, S5: And were any of these situations for personal reasons, business reasons or both? Base: All respondents — Total (1,600), Those who have a need to communicate in a foreign language (1,111)
The majority of the communication is taking place face-to-face with most using free help for their translation needs.

### Current communication habits

<table>
<thead>
<tr>
<th>Method</th>
<th>Business</th>
<th>Personal</th>
<th>Free</th>
<th>Paid</th>
</tr>
</thead>
<tbody>
<tr>
<td>Face-to-face</td>
<td>61%</td>
<td>70%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Telephone</td>
<td>34%</td>
<td>29%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Email</td>
<td>26%</td>
<td>24%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Text message/ SMS</td>
<td>13%</td>
<td>15%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Video call</td>
<td>7%</td>
<td>10%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Instant messaging</td>
<td>8%</td>
<td>8%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>None – I wanted to but wasn’t able to</td>
<td>19%</td>
<td>16%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Source: A2: How often have you needed or wanted to communicate in another language? A3: How have you communicated in another language?
Base: All respondents – Those who have a need to communicate in a foreign language (1,111)
Remote Conversations

1:1 conversation
2 languages
2 devices
Skype call with a friend

Brief exchange

1:1 conversation
1 device
Ordering food in Beijing

Extended social conversations

Many : many conversation
Dinner with extended family

Unidirectional Conversations

1: many conversation
Classroom or Lecture
Remote Conversations

1:1 conversation
2 languages
2 devices
Skype call with a friend

Brief exchange

1:1 conversation
1 device
Ordering food in Beijing

Extended social conversations

Many : many conversation
Dinner with extended family

Unidirectional Conversations

1: many conversation
Classroom or Lecture
Both active conversation and history include both languages. One will be dominant depending on who is speaking.
Both active conversation and history include both languages. One will be dominant depending on who is speaking.
When translated audio finishes playing, app listens for translation.
User feedback received on that design

“Skype Translator is also deaf to the rhythms of normal spoken conversation, so you can’t be quite sure when its disembodied robot voice is going to break in and start blurting out its translated version.”

"I know that this is a monumental task and will revolutionize technology... but there isn’t a flow in communication ..."
## Skype – defined conversation cadence

<table>
<thead>
<tr>
<th>Participant A</th>
<th>Translation</th>
<th>Participant B</th>
<th>Translation</th>
<th>Participant A</th>
<th>Translation</th>
<th>B</th>
<th>T</th>
</tr>
</thead>
</table>

## Natural Human Conversation

<table>
<thead>
<tr>
<th>Participant A</th>
<th>Participant B</th>
<th>Participant A</th>
<th>Participant B</th>
<th>Participant A</th>
</tr>
</thead>
</table>
How at
How are you fell
How are you feeling after
How are you feeling after surgery?
Hello!
Hola!
Cómo estás?
How are you?
Aber der große Unterschied ist die Art und Weise dieser Technologie ist Einsatz sozialer Medien, um den Wortschatz zu erweitern.

But the big difference is the way this technology is using social media to widen the vocabulary.
Ducking: Varying the volume of the original audio

Speaker hears the other person ducked during interpretation. Speaker hears his own translation always at low volume.
Text and Audio User Preference

a. Most Preferred Interface Condition in the Third Round

<table>
<thead>
<tr>
<th>Language</th>
<th>CC-TTS</th>
<th>CC</th>
<th>TTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fre-Ger</td>
<td>56%</td>
<td>33%</td>
<td>11%</td>
</tr>
<tr>
<td>Eng-Ger</td>
<td>63%</td>
<td>8%</td>
<td>38%</td>
</tr>
</tbody>
</table>

b. Least Preferred Interface Condition in the Third Round

<table>
<thead>
<tr>
<th>Language</th>
<th>CC-TTS</th>
<th>CC</th>
<th>TTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fre-Ger</td>
<td>14%</td>
<td>43%</td>
<td>43%</td>
</tr>
<tr>
<td>Eng-Ger</td>
<td>25%</td>
<td>75%</td>
<td>9%</td>
</tr>
</tbody>
</table>

Key

- Closed captions and translated audio
- Only closed captions
- Only translated audio
Typical lifetime of an utterance

User speaks → User reads partial transcript → User reads final transcript → Both hear final translation → Partner reads partial translation → Partner reads final translation → Partner speaks
Typical lifetime of an utterance

User speaks

User reads partial transcript

User reads final transcript

Both hear final translation

Partner reads partial translation

Partner reads final translation

Partner speaks
Repeat or rephrase

1. User speaks
2. User reads partial transcript
3. User reads final transcript
4. User realizes error

Repeat
Rephrase [ partial | complete ]
Type [ partial | complete ]
This is my transcript

This is my partner’s translated transcript
This is my transcript
This is my transcript

This is my partner’s translated transcript

Users don’t discover that they can type here
Where are we going?
MT quality

• Only one direction: up

• Huge quality differences between languages and content types

• Relatively slow progress

• Neural Networks provide a qualitative jump
  • Especially for languages with very different sentence structures
  • Bringing Chinese<>English up to similar quality as Spanish<>French
Machine Interpretation

• Starting to become useful for consumers
• Limited business application
• As a help for human interpreters
Speech Translator on Github

The web service API is publicly available.