Speech translation user experience in practice
Building a speech translation feature for Skype

Chris Wendt christw@microsoft.com
Will Lewis wilewis@microsoft.com
Tanvi Surti tsurti@microsoft.com
Microsoft Research, Redmond, Washington, USA

Abstract

Human conversations, especially between humans who are untrained in interpreted conversations, do not fit an utterance-translation-utterance-translation temporal sequence. Speech, especially speech by a human to another human, is burdened with many artefacts that are undesirable in translation. As implementers, we have a number of methods available to us, to alleviate these artefacts and provide a text to the translation system that looks more like written text than spoken text. We'll discuss the state and the limits of that approach, and the challenges on the journey of providing a representation of the speaker's intent.

1. Introduction

Skype is in the business of breaking down barriers to long-distance communication, crossing geographical distance. Automatic translation extends the concept to language, breaking down those barriers faster and cheaper, opening up more and more scenarios where communication across language boun¬daries becomes possible.

In the Star Trek TV series, the Universal Translator appears to be such an ordinary device that it almost seems surprising that translation of speech is actually new. Since its inception in 2003, Skype has been a primary vehicle for long-distance commu¬ni¬cation that easily bridges continents. Automatic translation of speech in a human-to-human fashion using a conversational domain is now extending the concept of removing barriers to communication to language.

2. Three Factors Coming Together

Why introduce wide domain conversational speech trans-lation now? Technology for recognizing speech has been around since the 1960s. Machine translation exists since the late 1950s. Putting them both together produced a multiplication of each other's mistakes, resulting in useless garbage. Speech translation is difficult because a single misrecognized word in the original language easily makes the target translation completely unusable. Looking at the output of automatic speech recognition (ASR) in the same language as spoken, mistakes don't appear so bad: the recognizer is likely to produce a homo¬phone of the intended word or phrase, and, with some imagination, a human can revert the misrecognized word back to the similar-sounding original, and make sense of the whole. The translation system, as designed for text translation, cannot recover from homophones, because after translation of the text, the words in the target language are not phonetically related to the correct alternative anymore. Once translated, the homophones don't sound alike at all anymore – the translation becomes incomprehensible.

Like speech recognition, deep neural networks were origin-ally proposed in the 1950s as well. But, due to the lack of processing power, they didn't show their revolutionary ability until recently. Over the last 10 years, deep neural networks have been successfully applied in speech recognition, in OCR and in image recognition. In speech recognition the introduction of deep neural networks gave us a 33% quality boost, reducing the word error rate from 30% to 20%, on an unbound domain English test set, with varying recording quality. A relative 33% reduction of the error rate is a truly drastic, enormous improvement [5], [6].

Automatic translation of text already had a major breakthrough in terms of quality and language coverage, within the last 5-10 years, due to the advances brought to the field by statistical machine translation (MT). Statistical MT gave us wide domain coverage and easy trainability for any language pair that has enough bilingual training material, and is producing good results in translating professionally or casually authored text in the major language pairs. Deep neural networks are on the verge of bringing another big boost in text translation, most visible in language pairs that have significantly different language structure, for instance when translating between English and Japanese.

Skype is the primary network for people wanting to connect across the earth's borders, oceans and continents, using their voice and video. This made Skype the perfect medium to help these people communicate across language barriers as well.

Three factors have now come together to create the opportunity for developing a speech translation system for conversational speech in an unbound domain: 33% quality improvement from deep neural networks for speech, statistical MT being reasonably good, and the border-crossing network of Skype.

3. Human-to-human Conversation

Our mobile devices and gaming consoles, our cars, our banks, and Siri, Cortana, Google Now, act on voice commands quite well, even complex commands and queries involving names of people or locations. With digital assistants, users are talking to a machine with a clearly limited scope at any given point in time, not another person. The systems are trained and optimized to perform well in a restricted scenario. A conversation between two humans doesn't follow any rules or defined vocabulary. It is unbound in terms of what the conversations participants say and how they say it. The domain of what the participants are talking about is unrestricted.

This of course poses a challenge. The larger the vocabulary, the more inherent acoustic ambiguities are present in the vocabulary, as close and not-so-close homophones, which drastically increase the set of alternatives the system needs to consider. When you say the word "wreck", for instance, the acoustic signal doesn't know word boundaries, so "wreck" can be the first part of "rec-ognize" or the last or middle part of some other words. The acoustic model and the language models have consumed many thousands of hours of audio and transcripts from real people talking to each other, in order to build their knowledge of context, and the suitable alternatives within the context.

Deep neural networks in speech recognition, due to their depth, can consume a wide variety of accented speech, and different ways to say a word, without degrading the experience for other accents. They are better suited to remember the context as seen during training, and can reproduce the context at runtime. As a result, Skype doesn't need different settings for English spoken on different continents – it uses the same large models that are trained with British, Canadian, American, Australian and New Zealand speakers, together. The same is true for French and Spanish. The resulting system may still do better with one accent or another accent,

but that's only due to the amount and acoustic variations within the training material. The training cycle can add this material as it becomes available, without degrading the experience for anyone.

4. Intent vs. Received Audio

When humans speak to each other, they are typically not very cognizant of grammar, fluency, clarity, casing and punctuation, far less so than when putting down the written word.

What he thought he said:

Yeah. I guess it was worth it.

What he actually said:

Yeah, but um, but it was you know, it was, I guess, it was worth it.

When the MT system translates what he really said, it doesn't get better or easier to understand.

Microsoft uses a component called TrueText [1] to remove disfluencies and other undesired artefacts of human speech: Grunts and coughs, ahs and ems, false starts and repeats. TrueText is a translation system by itself – it translates spoken language to a series of tokens that look more like written language within the same language. Written language is what the actual translation system is good at. TrueText is trained on pairs of exact transcripts and fluent transcripts of the same spoken utterance. A typical human transcript is in fact a fluent, clean transcript. The exact transcript can be synthesized from the original audio and the fluent transcript, thereby creating the parallel training data for the TrueText translation system.

5. Design Considerations

5.1. Interaction with Humans

Does an interpreter have a persona? Should the interpreter have a persona, or rather be transparent, in the background, not an actor in the conversation? We can observe that a novice at interpreted conversations tends to perceive the interpreter as an actor, a person who can be addressed directly, maybe even as a cultural consultant. Experienced participants in interpreted conversations will remain focused on the foreign-language conversation partner, and let the interpreter be the medium, the conveyor of information, visible and audible, but not interfering in the conversation. In settings like the European Parliament, this understanding of the hidden interpreter is helped by the fact that the interpreter is in fact far away and invisible.

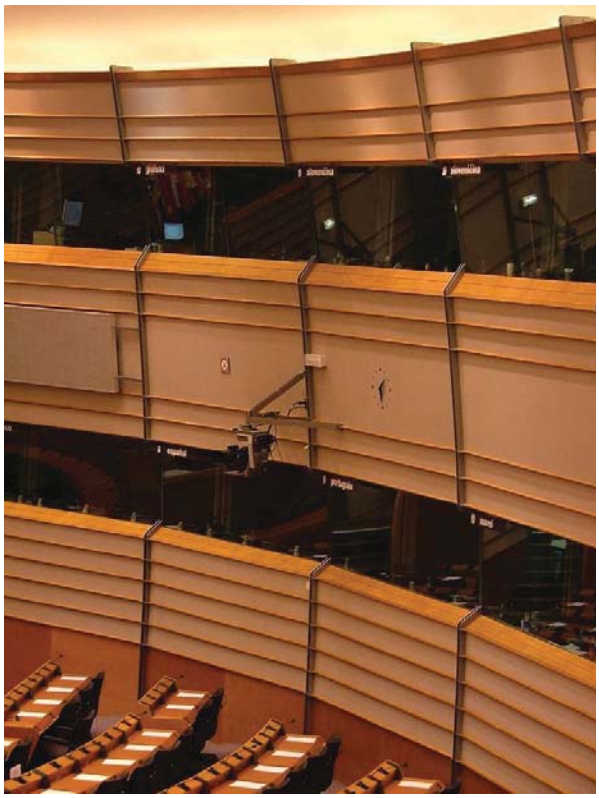

Figure 1. At the European Parliament, the interpreters are mostly invisible

With Skype Translator, Microsoft had a choice to model the translation function with an acting persona, or transparent like the EP interpreter. The acting persona implementation has the psychological advantage that there is someone to blame for misinterpretation and misunderstanding. "No, I didn't say that, it is the interpreter's fault. ". If the system remains persona-less, the utterance is more directly attributable to the person who spoke the original, even if it was the recognition or translation that introduced the mistake. The Skype Translator team decided to go persona-less, and just put the conversations between the humans in the foreground. It is not only easier to implement in a functional fashion, it is also more conducive to productive conversations.

Figure 2. Chinese President Xi Jinping visits Microsoft, with Microsoft CEO Satya Nadella and interpreter

Maybe there is a time when the interpreter can get a persona and interact with the participants, in a fun and entertaining way.

5.2. Handling the Audio Stream

In the physical world, an interpreted conversation goes something like this:

- A speaks, and then A stops speaking.
- Interpreter speaks what A just said, in B's language, and then stops speaking.
- B speaks, and then B stops speaking.
- Interpreter speaks what B just said, in A's language, and then stops speaking.

That is a fairly long and drawn-out process, in fact twice as long as a direct, same-language conversation. It is complicated by the fact that B won't know when A's interpretation is finished. B doesn't know what A said, so B won't know the logical end of the utterance, there would need to be some kind of signal indicating the end of the utterance, so that B knows when to start speaking. Humans can use gestures and facial expressions to indicate when they are done.

Skype can do away with all this, making use of the fact that Skype controls what each participant hears. After all, they are on Skype, and not in the same room. Skype uses a broadcasting technique called "ducking", which ducks the volume of the original voice under the voice of the interpreter. This is commonly used in broadcasts of interviews with foreign dignitaries: First you hear a few words of the foreign dignitary speaking in the foreign language at full volume, then the interpreter kicks in at full volume, and the volume of the dignitary's voice becomes low enough to duck under the interpreter's voice. Skype does the same: You hear your

partner at full volume first, then the interpretation kicks in at full volume. Your partner can continue to speak and you'll hear her voice ducked under the interpreter's voice, as long as the interpretation goes on, and then it comes back at full volume.

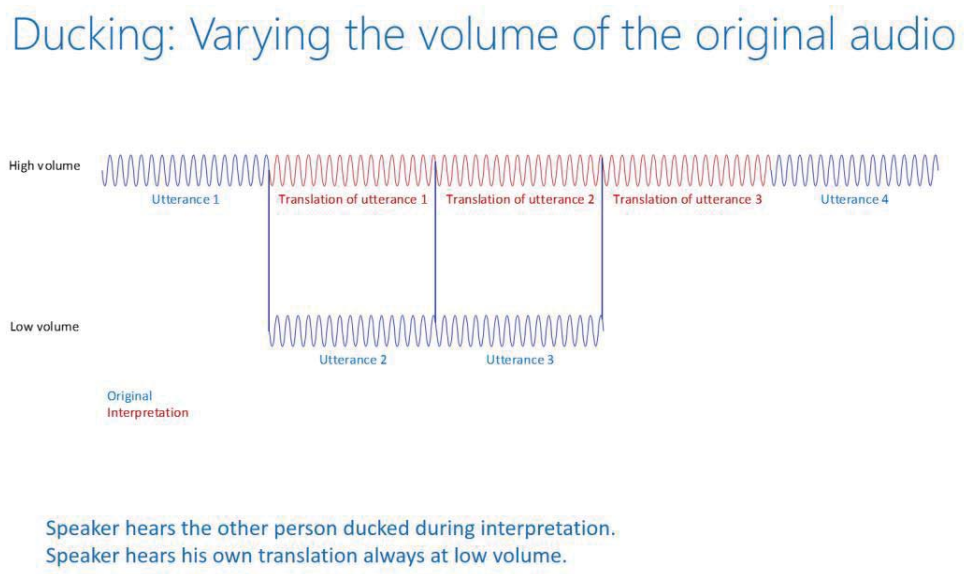

Figure 3 Ducking: Varying the volume of the original audio

You will hear the interpretation of what you yourself said, ducked under everything, always faint. It shows that the system is working, explains why your partner didn't laugh yet about your super funny joke – she didn't hear it yet. Skype could have chosen a different indication that "interpretation of what you just said is in progress", but the team figured the actual ducked audio of your interpretation is the best indicator. Speaking over your own interpretation is entirely possible and liberating: you can tell the whole long story of your weekend adventures, without having to pause.

In addition to the translated audio, Skype also shows you a transcript of what you said, and the translation to your language of what your partner said. In the usability tests [3], about half of the participants preferred to turn the translated audio off and just read the transcripts. Indeed, it makes for a more fluent conversation – most of us are faster reading than we are listening. However, the translated audio works better on small-factor devices and in situations where you don't want to, or can't look at the screen the whole time.

6. Observations in Practice

6.1. Named Entities

While Skype Translator is designed with a large vocabulary, allowing you to speak about anything and everything, the vocabulary does have limits, which is most painful in the area of named entities. Names of places, people or businesses fail to be recognized correctly, in the majority of cases. With a few exceptions: at the start of the call, Skype builds a mini-language model from your personal Skype address book, and adds it to the vocabulary, in the same way your phone does, to allow you to call the people in your address book by voice command. A

difficult problem to overcome is that these added names are synthesized in the pronunciation of your language. If you are an English speaker, you'll have to pronounce all names in an anglicized manner, for the people's names to be recognized. That's probably not how you would properly pronounce foreign names in your address book, which most likely happen to be the ones you will be mentioning during a translated call. Adding the entire set of the world's first and last names, all place names and all business names to the vocabulary would degrade the accuracy so much, the system would become unusable if that happened. For the foreseeable future the system will need to employ selective methods of dynamically adding just the right amount of phrases to the vocabulary, for the given situation, by taking intent and location into account.

6.2. Accented Speech

Skype Translator's training material mostly comes from native speakers of their language. The audio characteristics of second language speakers significantly differ, and second language speakers will find a noticeable drop in recognition quality, compared to native speakers.

6.3. Audio Quality

Speech recognition in a wide domain requires a microphone that sits close to your mouth, in order to pick up the nuances of your voice and to reduce the impact of background noise. Such a microphone comes with a headset that also gives you better audio for your own listening. Skype users are used to having Skype conversations simply using the laptop's or tablet's built-in microphone and speakers.

6.4. Pronunciation

Practicing better pronunciation helps a lot. Some users expressed that using Skype translator helps them to become better native speakers – forcing them to exercise more careful pronunciation. It also helps to know what you want to say before you say it – a feat that the non-politicians among us sometimes have trouble with.

7. Conclusion

Are we done with automatic interpretation? Not by far. There is lots of room for growth: Recognition accuracy and spoken language artefact removal, while maintaining and growing the very wide domain coverage, will get better drastically.

Translation itself benefits from the introduction of neural networks, which have introduced a qualitative jump in other areas of machine learning. Neural networks have the ability to remember any number of factors that influence a particular translation, much better than today's statistical systems do. That helps translating between languages of different language structure as well as lifting the quality of the lower-resourced languages, making better use of smaller amounts of parallel material.

User experience benefits from adjustment to the usage scenario: While the experience in Skype benefits the use in long-distance video calls, it will evolve for in-person meetings or group calls.

Microsoft makes the API that powers translation in Skype [4] available for your own communication applications as well. You can use it to build close-to-real-time speech recognition and translation solutions for your own scenarios.

References

This paper cites large sections from [2] in unchanged form.

[1] Lee Schwartz, Dilek Hakkani-Tür, Gokhan Tur, Hany Hassan Awadalla, "Segmentation and Disfluency Removal for Conversational Speech Translation" Proceedings of Interspeech, ISCA - International Speech Communication Association, September 1, 2014.

[2] Chris Wendt, "Behind Skype's machine interpreting", Multilingual Magazine, September 2016, https://multilingual.com/all-articles/?art_id=2373

[3] Kotaro Hara, Shamsi Iqbal, "Effect of Machine Translation in Interlingual Conversation: Lessons from a Formative Study", ACM Conference on Human Factors for Computing Systems (CHI), April 1, 2015.

[4] Microsoft Corporation, "Microsoft Translator", http://www.microsoft.com/translator.

[5] Li Deng, Jinyu Li, Jui-Ting Huang, Kaisheng Yao, Dong Yu, Frank Seide, Michael Seltzer, Geoff Zweig, Xiaodong He, Jason Williams, Yifan Gong, Alex Acero, Mike Seltzer, "Recent Advances in Deep Learning for Speech Research at Microsoft", IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), May 1, 2013.

[6] Dong Yu, Frank Seide, Gang Li, "Conversational Speech Transcription Using Context-Dependent Deep Neural Networks", ICML 2012, June 1, 2012.