Microsoft Speech Language Translation (MSLT) Corpus: The IWSLT 2016 release for English, French and German

Christian Federmann, William D. Lewis

Microsoft Translator
Microsoft Research, Redmond, WA, USA

{chrife,wilewis}@microsoft.com

Abstract

We describe the Microsoft Speech Language Translation (MSLT) corpus, which was created in order to evaluate end-to-end conversational speech translation quality. The corpus was created from actual conversations over Skype, and we provide details on the recording setup and the different layers of associated text data. The corpus release includes Test and Dev sets with reference transcripts for speech recognition. Additionally, cleaned up transcripts and reference translations are available for evaluation of machine translation quality. The IWSLT 2016 release described here includes the source audio, raw transcripts, cleaned up transcripts, and translations to or from English for both French and German.

1. Introduction

We describe the Microsoft Speech Language Translation (MSLT) corpus that we created to evaluate end-to-end quality of Microsoft Translator's speech translation system, the system powering speech translation in Skype Translator. The goal of Skype Translator is to support open-domain, spontaneous conversations between individuals who speak different languages, ultimately such that one would not be able to tell the difference (*e.g.*, semantically or in latency) between conversations held in one language and those held in two. We carefully constructed the corpus so that the Test and Dev sets would not be constrained by the current state-of-the-art; in effect, we wanted our test data to represent the gold standard of our ultimate aspirations for speech language translation.

The MSLT corpus currently contains full end-to-end speech translation Test and Dev sets for three languages: English, French and German. To adequately test the conversational translation scenario, and any components used in speech translation, *e.g.*, Automatic Speech Recognition (ASR), disfluency processing, Machine Translation (MT), etc., whether combined or in isolation, the MSLT corpus consists of the following data: audio files, their verbatim transcriptions (*i.e.*, faithfully capturing the speech input, including disfluencies), cleaned-up transcripts (*i.e.*, where disfluencies are removed), and translations. The corpus contains subcorpora for each pair in and out of English (so, English→French, French→English, English→German, German→English). This paper presents details on the recording, transcription and translation processes we conducted to build this corpus.

2. Data Collection

2.1. Recording Modes

For our experiments with speech translation we attempted to create realistic Test and Dev sets. Crucially, we wanted test data that represented actual bilingual conversations, and was not constrained by the current state-of-the-art or domain.[1] Ultimately, it is our research aspiration to develop open-domain speech translation that differs little with respect to domain, latency, etc., from monolingual conversations or from bilingual conversations between fluent bilinguals. Thus, we wanted test data that came as close as possible to fully bilingual conversations, or fluently translated monolingual conversations. Initially, we paired consultants and tested several recording scenarios in a limited study, comparing different *modes* of conversations between two speakers:

1. Monolingual conversations between English speakers
2. Bilingual conversations between English and Spanish speakers with an automated speech translation module (*i.e.*, using ASR)
3. Bilingual conversations without a translation module between bilingual speakers

The first scenario allowed us to identify basic properties of conversations which are not subject to any translation ef-

[1]The FISHER corpora, *e.g.*, for English, (LDC2004T19 and LDC2005T19[1, 2], for Spanish, LDC2010S01 and LDC2010T04[3, 4], as well as the CALLHOME corpora[5, 6, 7, 8] are similar to the MSLT, in that the audio consists of free-form phone conversations. There are two differences: (1) FISHER and CALLHOME contain recordings of low-bandwidth phone audio data. Given that Skype supports higher bandwidth signals, we wanted test data that was more representative of that scenario. (2) FISHER and CALLHOME were designed for testing ASR, and do not contain full end-to-end S2S content, *i.e.*, no cleaned-up transcripts and no reference translations for machine translation.

fects. In fact, we ran two such experiments, one for English, the other for Spanish. We found no significant differences between the two languages in our study.

The second scenario tested speech-to-speech (S2S) translation effects in real life. We observed a clear negative impact of the translation module, most notably that speech rate dropped compared to the monolingual scenarios. Also, vocabulary was more constrained. Both effects can be attributed to quality problems with the speech translation module used for our experiment, and would likewise exist with any state-of-the-art speech translation system. Since no current implementation of speech translation can reach perfect human quality, noisy and imperfect output from an S2S engine increases the need to repeat utterances that were not transcribed correctly (*i.e.*, the ASR failed to recognize the utterance), or rephrase utterances that were not understood (*i.e.*, not translated correctly). Further, users will often need to ask clarification questions when results are not understandable. All of this slows down the conversation and impacts its flow.

The third scenario dropped the translation module. Instead, we worked with fluent bilingual speakers, who would naturally understand utterances in either language. In this scenario, one speaker spoke in one language, the other speaker in the other language. Scenario three is a closer approximation of our aspired goal, in that speakers are actively engaged in fluent bilingual conversations (granted, without a speech translation module). In our experiments, this scenario proved to be a good compromise between recording quantity, quality and applicability to our goal of speech translation evaluation.

2.2. Recording Guidelines

We ultimately opted for the third recording mode and recorded conversations between bilingual speakers. Although the second scenario may, on the surface, most closely resemble our eventual use case (*e.g.*, bilingual translated calls over Skype), by using speech translation in the study, we realized that we would be constraining our test data to the current state-of-the-art. This would limit the utility and long-term viability of our test data.

For the recordings, there was no translation module involved. Recordings were conducted using a specially designed preview version of Skype Translator with translation turned off (in other words, only the audio was being captured, but no translation was provided). Using Skype Translator to record the conversations allowed us to capture typical side effects of Skype's transport layer. This makes our data more realistic compared to engineered data recorded under optimal conditions.

For each pair of speakers, we organized conversations as follows:

- Speakers recorded two sessions, 30 minutes each
- Speakers switched roles: one spoke the native language in one conversation, English in the other
- Conversations are lightly constrained to predefined topics (topics were used more to prime conversations than to act as constraints)

We recorded at least 100 speakers for each language, with 50+ pairings. Speakers were balanced for gender and age groups. The English side of the recordings for French and German bilingual conversations were discarded as they represent accented speech. For English, we collected data from monolingual English conversations between speakers of different English dialects (American, Australian, British and Indian), ensuring speaker and dialect diversity.

2.3. Annotation Guidelines

We asked annotators to transcribe the given audio signal in disfluent, verbatim form. Incomplete utterances and other sounds are transcribed using:

- predefined **tags** such as <SPN/> or <LM/>, and
- free **annotations** such as [laughter] or [door slams].

In theory, annotators are free to choose whatever annotations they deemed appropriate for sounds which none of the predefined tags captured. In reality we observed only one such annotation: [laughter].

The following list provides details on the predefined tags and their interpretation.

- **SPN: Speech noise:** Any sounds generated during speaking which are not actual words should be transcribed as speech noise. Examples are lip smacks or breathing noises from the primary speaker.
- **EU: End unspoken:** Used when the end of a word was truncated or swallowed by the speaker, possibly due to hesitation. Example: "hell<EU/> hello".
- **NON: Non-speech noise:** Any sounds which are not generated by a speaker should be transcribed as non-speech noise. Examples are external sounds such as cars or music from a TV running in the background.
- **UNIN: Unintelligible:** When the transcriber cannot even make an educated guess at which word has been uttered by the speaker, it should be transcribed as unintelligible. Should be applied to one word at a time. For multiple such words, multiple tags should be used.
- **LM: Language mismatch:** If the word uttered by the speaker is understandable but not in the expected speech language the annotator should use the language mismatch tag. If the foreign word can be identified, it should be embedded into the tag, otherwise an empty tag is sufficient. Examples are "Hello <LM>monsieur</LM>" or "I visited <LM/>".

– **AS: Audio spill:** If the audio signal is polluted by feedback or audio bleeding from the second channel or affected by any other technical issues, this should be transcribed as audio spill. Generally, this indicates bad headsets or recording conditions.

– **SU: Start unspoken:** Used when the beginning of a word was truncated or otherwise messed up by the speaker. Example: "<SU/>an hear you".

– **UNSURE: Annotator unsure:** Indicates a word the transcriber is unsure of. Should be applied to one word at a time. For multiple such words, multiple tags should be used.

– **NPS: Non-primary speaker:** Indicates a word or phrase which has been uttered by a secondary speaker. This speaker does not have to be identified. Example: "watching the water flow. <NPS>yeah.</NPS>"

– **MP: Mispronounced:** A mispronounced but otherwise intelligible word. Example: "like, a filet <MP>mignon</MP>"

Table 2 gives a detailed overview on the observed frequencies of these tags for each of the released MSLT data sets.

3. Corpus Data

3.1. Audio Files

The corpus contains uncompressed WAV audio files with the following properties:

– **Encoding:** PCM
– **Sample rate:** 16,000 Hz
– **Channels:** 1, mono
– **Bitrate:** 256 kbit/s

Note that the original audio streams had been encoded using the Siren codec so we had to transcode them to create the uncompressed files for release. Furthermore, the original signal had been subject to transport via Skype's network with variable bandwidth encoding. Audio quality of the released files may be affected by both factors. Files represent a realistic snapshot of speech quality in real life.

Table 1 gives more details for the audio portions of the MSLT release.

Language	Data set	Files	Runtime	Average
English	Test	3,304	4h03m58s	4.4s
	Dev	3,052	3h56m37s	4.7s
French	Test	2,120	3h26m46s	5.8s
	Dev	2,381	3h36m30s	5.4s
German	Test	2,275	3h56m53s	6.2s
	Dev	2,074	3h50m29s	6.7s

Table 1: Audio runtime information for our Test and Dev data by source language.

3.2. Text Files

Transcripts (T1, T2) and translations (T3) are formatted as Unicode (16 bits, little-endian) text files. We defined these three text annotation layers for our speech-to-speech processing:

– **T1: Transcribe:** results in a raw, human transcript which includes all disfluencies, hesitations, restarts, and non-speech sounds. The goal of this annotation step is to produce a verbatim transcript which is as close to the original audio signal as possible. We observed bias when speakers annotated their own transcripts (repairing, *e.g.*, disfluencies and restarts), so we assigned work to a different set of consultants to prevent this issue. Both punctuation and case information are optional in T1 but we found that annotators already provided this.

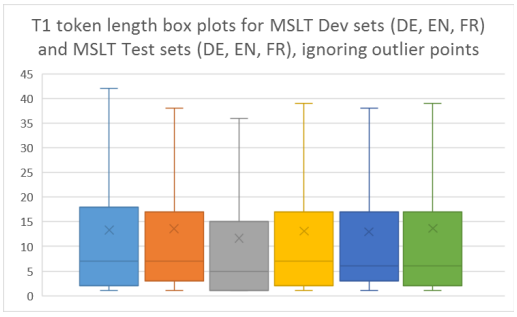
– **T2: Transform:** represents a cleaned up version of the T1 transcript with proper punctuation and case information. T2 output also should be segmented into *semantic units* and should not contain any disfluencies. Annotators work on the T1 text files only and do not have access to the original audio files. The idea is to create conversational text which might be printed in a newspaper quote. Segmentation and disfluency removal may introduce phrasal fragments, which are kept as long as they have at least *some* semantic value.

– **T3: Translate:** represents the translation of the fluent T2 transcript. The goal is to create conversational target text which feels natural to native speakers. Translations have been created based on unique segments in order to enforce translation consistency. Translators are instructed not to translate any (remaining or perceived) disfluencies but instead asked to flag such T2 instances for repair.

3.3. Corpus Statistics

Figure 1 shows box plots for all MSLT data sets. The left graph focuses on token length for disfluent T1 transcripts. We ignore outlier points. There are no significant differences between the sets, neither by type, nor by language. The vast majority (typically around 80% for each of the sets) of all transcripts has a token length smaller than 15.

Numbers for our T2 transcripts show expected behavior: segment counts increase and the token numbers decrease. The right graph shows consistently lower box plots for all data sets, on the same scale.



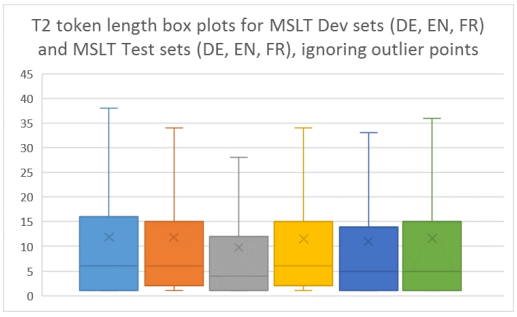


Figure 1: Token length box plots for German, English and French Dev and Test sets, ignoring outliers. Box plots from left to right: Dev DE, Dev EN, Dev FR, Test DE, Test EN, Test FR. Left graph shows T1 transcripts: there are no significant differences between languages or sets. Right graph shows T2 transcripts: token counts are generally lower than for verbatim T1 transcripts.

Table 3 provides an overview on segment, token and type counts for all both Test and Dev data for English, French and German. Note the significantly higher number of tokens for both English sets. A possible explanation lies in the fact that English conversations were easier as speakers only had to "translate" between different English dialects. Hence, these conversations were much closer to our monolingual recording scenario than conversations for French or German.

3.4. Example

Figure 2 gives an example containing disfluent, verbatim transcript (T1), cleaned up and transformed text (T2) and the corresponding translations (T3) into French and German. Note how T2 transformation breaks the T1 transcript into two segments and also removes the `[laughter]` annotation. Translations are aligned on the segment level.

4. Conclusion

We presented the Microsoft Speech Language Translation (MSLT) corpus for end-to-end evaluation of speech translation systems and/or component level evaluation. In the latter case, the MSLT Test data consists of component level data: to test the ASR component, MSLT has audio data and verbatim transcripts; to test disfluency removal and related processing, MSLT has transcripts that have been cleaned up of disfluencies, restarts, hesitations, laughter, and any other content not relevant to translation; to test conversational MT, MSLT has translated transcripts. Because speech-to-speech by its nature is bidirectional, test data for any language pair has the full pipeline for both directions.[2,3] While the initial release of MSLT was targeted at IWSLT 2016 participants, we intend to release an updated version in 2017, principally to expand language coverage, but also to make fixes to the existing data and recordings (as needed).

[2]It should be noted that the conversations recorded for either direction for any given language pair are not semantically contiguous, that is, they do not consist of recordings of the same conversation sessions. This is due to the fact the English side of French and German conversations was thrown out due to non-English accents, and that all kept English sessions were recorded separately.

[3]It should also be noted that the test data *assumes* disfluency processing, since the data that has be translated has been cleaned up. In other words, we assume in an S2S workflow that MT is handed cleaned content. [9] suggest an alternate workflow where the ASR output is not cleaned up, and an MT system is trained on noisy, ASR-like content. Our test data could be used to test the transcription of an S2S system built in such a way, but it could not be used to test the entire end-to-end workflow as we don't have translations of the disfluent transcripts.

5. References

[1] C. Cieri, D. Graff, O. Kimball, D. Miller, and K. Walker, "Fisher English Training Speech Part 1 Transcripts LDC2004T19," Web Download. Philadelphia: Linguistic Data Consortium, 2004. [Online]. Available: https://catalog.ldc.upenn.edu/LDC2004T19

[2] ——, "Fisher English Training Part 2, transcripts LDC2005T19," Web Download. Philadelphia: Linguistic Data Consortium, 2005. [Online]. Available: https://catalog.ldc.upenn.edu/LDC2005T19

[3] D. Graff, S. Huang, I. Cartagena, K. Walker, and C. Cieri, "Fisher Spanish Speech LDC2010S01," Web Download. Philadelphia: Linguistic Data Consortium, 2010. [Online]. Available: https://catalog.ldc.upenn.edu/LDC2010S01

[4] ——, "Fisher Spanish – Transcripts LDC2010T04," Web Download. Philadelphia: Linguistic Data Consortium, 2010. [Online]. Available: https://catalog.ldc.upenn.edu/LDC2010T04

[5] A. Canavan and G. Zipperlen, "CALLHOME Spanish Speech LDC96S35," Web Download. Philadelphia: Linguistic Data Consortium, 1996. [Online]. Available: https://catalog.ldc.upenn.edu/LDC96S35

[6] B. Wheatley, "CALLHOME Spanish Transcripts LDC96T17," Web Download. Philadelphia: Linguistic Data Consortium, 1996. [Online]. Available: https://catalog.ldc.upenn.edu/LDC96T17

[7] A. Canavan, D. Graff, and G. Zipperlen, "CALLHOME American English Speech LDC97S42," Web Download. Philadelphia: Linguistic Data Consortium, 1997. [Online]. Available: https://catalog.ldc.upenn.edu/LDC97S42

Annotation	Description	English		French		German	
		Test	Dev	Test	Dev	Test	Dev
<SPN/>	Speech noise	200	271	513	531	778	655
<EU/>	End unspoken	409	388	152	173	141	143
<NON/>	Non-speech noise	192	235	92	81	71	122
<UNIN/>	Unintelligible	306	125	111	101	73	73
<LM/>	Language mismatch	12	0	127	236	55	148
<AS/>	Audio spill	6	0	105	229	1	1
<SU/>	Start unspoken	37	54	64	51	22	30
<UNSURE/>	Annotator unsure	59	81	31	28	6	4
<NPS/>	Non-primary speaker	44	68	13	19	4	1
<MP/>	Mispronounced	3	4	5	4	1	6
[laughter]	Laughter	217	192	228	308	194	226
	Annotations	1,487	1,418	1,441	1,761	1,346	1,409
	Utterances	3,304	3,052	2,120	2,381	2,275	2,074
	Tokens	42,852	41,450	28,926	27,749	29,903	27,688
	Types	36,318	35,308	23,114	22,646	25,523	23,768

Table 2: Annotation information for our Test and Dev data by source language.

Language	Type	Segments	Tokens	Types
English	T1 (EN)	3,304	42,852	36,318
	T2 (EN)	5,175	36,388	31,981
	T3 (DE)	5,175	37,324	33,862
	T3 (FR)	5,175	39,776	35,614
French	T1 (FR)	2,120	28,926	23,114
	T2 (FR)	3,602	24,728	21,409
	T3 (EN)	3,602	24,642	20,987
German	T1 (DE)	2,275	29,903	25,523
	T2 (DE)	3,928	26,247	23,633
	T3 (EN)	3,928	26,595	23,838

Language	Type	Segments	Tokens	Types
English	T1 (EN)	3,052	41,450	35,308
	T2 (EN)	5,313	36,184	31,960
	T3 (DE)	5,313	36,409	32,913
	T3 (FR)	5,313	40,159	35,824
French	T1 (FR)	2,381	27,749	22,646
	T2 (FR)	3,939	23,383	20,541
	T3 (EN)	3,939	23,596	20,370
German	T1 (DE)	2,074	27,688	23,768
	T2 (DE)	3,529	24,639	22,037
	T3 (EN)	3,529	25,077	22,260

Table 3: Segments, tokens and types for our Test and Dev data by source language and annotation type.

[8] P. Kingsbury, S. Strassel, C. McLemore, and R. McIntyre, "CALLHOME American English Transcripts LDC97T14," Web Download. Philadelphia: Linguistic Data Consortium, 1997. [Online]. Available: https://catalog.ldc.upenn.edu/LDC97T14

[9] G. Kumar, M. Post, D. Povey, and S. Khudanpur, "Some insights from translating conversational telephone speech," in *Proceedings of ICASSP*, Florence, Italy, May 2014. [Online]. Available: http://cs.jhu.edu/ gkumar/papers/kumar2014some.pdf

Language	Type	Segment	Text
English	T1	1	no, no, Bernie, Bernie is a Democrat. `[laughter]` Bernie is a socialist.
	T2	1	No, Bernie, Bernie is a Democrat.
		2	Bernie is a socialist.
French	T3	1	Non, Bernie, Bernie est un démocrate.
		2	Bernie est un socialiste.
German	T3	1	Nein, Bernie, Bernie ist ein Demokrat.
		2	Bernie ist ein Sozialist.

Figure 2: Example showing different layers of text annotation for an English utterance.