The Microsoft Speech Language Translation (MSLT) Corpus for Chinese and Japanese: Conversational Test data for Machine Translation and Speech Recognition

Christian Federmann chrife@microsoft.com
William D. Lewis wilewis@microsoft.com
Microsoft Translator
Microsoft AI+Research, Redmond, WA, USA

Abstract

Recent years have seen unprecedented growth in the use of MT across industries and domains. Partly this is due to the ready availability of open source MT tools such as Moses or online or customizable services. It is also due to fundamental shifts in the technology, specifically the move to deep learning, which has dramatically improved the quality of MT engines, including those used by online services. Likewise, improvements in Speech Recognition (SR) technology, also driven by the move to deep learning, are showing significant improvements in quality driven by deep learning alone. The improvements of both of these technologies, MT and SR, increase the potential viability for speech translation, since the error cascade caused by daisy-chaining these technologies drops as the quality bar raises. MT is a crucial component in speech translation systems, yet developing conversational MT systems essential to speech translation is not a focus for many working in the Machine Translation discipline. Particularly problematic for many languages is the absence of test and dev data, not any less true for the Chinese and Japanese languages, where forays into conversational MT in and out of these languages are limited by the lack of publicly available conversational test data. In this paper, we seek to address this problem, by providing MT test and dev data that has been built from actual bilingual conversations between English and Japanese and Chinese, test data that can be useful to drive further research in this space for these two languages. Our plan is to make the data described in this paper available to the public by MT Summit.

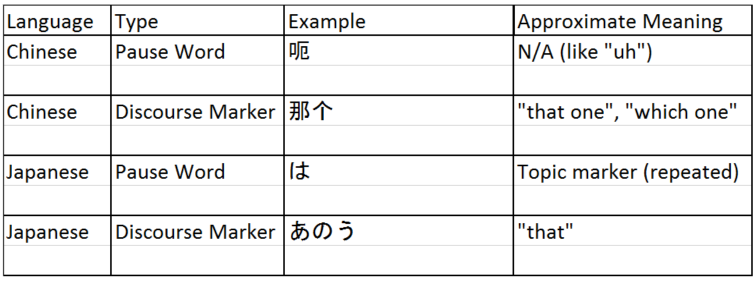
1 Introduction

The commoditization of MT, as evidenced by increased use of MT across industries and domains, results most significantly by the availability of open source MT tools such as Moses [1] and online tools for training and building customized systems (such as those offered by Microsoft, SDL, IBM, etc.). But it is also due to fundamental shifts in the technology, specifically the move to deep learning, which has dramatically improved the quality of MT engines [2, 3, 4], including those used by online services (*e.g.*, Google, Microsoft, Baidu, etc.). Likewise, improvements in speech recognition technology, also driven by the move to deep learning, are showing 25-50% improvements in quality driven by deep learning alone; for instance, [5] showed a 32% reduction in Word Error Rate when switching from Gaussian Mixture Models (GMMs) to Deep Neural Networks (DNNs), *with no change in training data*. The improvements of both of these technologies increase the potential viability for speech translation, since the error cascade caused by daisy-chaining these technologies drops as the quality bar in each component technology increases.

MT is a crucial component in speech translation systems, yet developing conversational MT systems essential to speech translation is not a focus for many working in the Machine Translation discipline. With the increase of conversation-like sources that needed to be translated, however, *e.g.*, social media, and an increase in the availability of speech recognition systems across multiple languages, translating less formal content is becoming far more commonplace and in-demand. MT systems that are trained on more "general" content, say, Web page content or parallel PDF documents, do not do well on content of a radically different style [6, 7].

Also problematic for developing conversational MT systems is an adequate way to evaluate them. Our focus in this paper is on test data for Chinese and Japanese, in and out of English. We describe here test and dev data that has been built from actual bilingual conversations between English to and from Japanese and Chinese. This test data consists not only of the audio (not relevant to MT *per se*, but certainly to speech translation), but also "raw" un-edited transcripts of the audio, cleaned-up caption-like transcripts, and translations to/from English and Japanese and Chinese. For each language there are two test sets: one from English, and one to English. This provides data that is native in the source language, eliminating problems with direction-bias in evaluation. Also, because the data is conversational, it is fully appropriate to tune systems to, and evaluate systems on, conversational-style content.

It should be noted that the bilingual English↔Japanese and English↔Chinese conversations were unscripted; participants were given some guidance with respect to topic, but otherwise, were allowed to have unrestricted conversations with one another. In many ways this is similar to the instructions provided to participants in the construction of the monolingual Switchboard corpus [8]. Because the conversations were unscripted, the data is rife with content typical of conversations: filler pause-words (um, uh), discourse markers (you know, I mean), restarts (I'm...I've), stutters (I-I-I), colloquial forms (gonna, kinda), and a host of disfluencies and artifacts common to colloquial speech. See Figure 1 for a few examples of Chinese and Japanese disfluencies. This provides a means to test MT systems designed for less formal, more conversation-like content, not only limited to output from speech recognition, but also content common in social media. Although our test data does not translate disfluencies *a la* [9], which was by design, it does preserve the informal character of the input content in the translation.



Language	Type	Example	Approximate Meaning
Chinese	Pause Word	呃	N/A (like "uh")
Chinese	Discourse Marker	那个	"that one", "which one"
Japanese	Pause Word	は	Topic marker (repeated)
Japanese	Discourse Marker	あのう	"that"

Figure 1: Example disfluencies in Chinese and Japanese.

2 Data Collection

The focus of our work here was to create realistic test data for evaluating conversational MT systems. Thus, we wanted the test data that reflected actual *bilingual* conversations between fluent speakers of English, Chinese and Japanese. Monolingual corpora of this type exist, *e.g.*, LDC2004T19 and LDC2005T19 [10, 11] for English, and the CALLHOME corpora for English and Chinese and other languages [12, 13, 14, 15] (but notably, *not* Japanese). The focus of these corpora, however, are explicitly on Speech Recognition and not Machine Translation, since no translations of the content are publicly available. For those interested in conversational Machine Translation or Speech Translation, these corpora are of little use, unless one wishes to expend significant resources in translation. Further, one has no options from Japanese, since these corpora do not cover Japanese.[1]

The "realistic" test data requirement also meant, crucially, that we did not want our test data to be constrained by the current state of the art in speech recognition and machine translation technology, nor constrained by domain. This requirement meant that we avoided using existing speech recognition and machine translation systems in data collection.

2.1 Recording Guidelines

As noted, we opted not to use existing speech recognition or machine translation technology in our recordings. This was motivated, in part, by experiments we conducted on English, German and French [17]. In these experiments, we noted that users behavior changed dramatically when having conversations mediated by a speech translation system: speech rate dropped dramatically from monolingual conversations, vocabulary was more constrained, conversations were punctuated by a significant number of restarts and rephrases, and users would often ask questions solely for the purpose of clarification (*e.g.*, when ASR or MT failed), affects that would not be present in fluent conversations. *Our interest is in constructing **realistic** bilingual conversations*, notably not constrained by the current state of the art, specifically sans the artifacts described above. Constraining systems in such a way would set a ceiling to what exists currently, and thus not provide a true gold standard conversational content to evaluate against.

[1] The BTEC corpus [16] does contain quasi-conversational Japanese input, but it is focused on the travel domain, and does not consist of free form conversations and transcripts.

Without the aid of machine translation, we could only recruit bilingual speakers of English, Chinese and Japanese. Bilinguality varies significantly across speakers, but generally, speakers will be dominant in one of the languages (usually their native or mother tongue) and less capable in other language(s). We recruited fluent bilingual speakers, who would naturally understand utterances in either language, but for the source language, only those who were dominant in that language. Thus, Japanese bilingual speakers needed to be native (or dominant) in Japanese, but capable of understanding English. Likewise for Chinese speakers.

For the recordings, no machine translation was used. Users held conversations over communication software installed on their computers, specifically Skype, and we captured the audio from these conversations. The audio data was then segmented into smaller chunks (typically, less than 30 seconds long) and transcribed faithfully, capturing all disfluencies present in the audio signal.[2] For each pair of speakers, we organized recordings adhering to the following paradigm:

- Speakers recorded two sessions, 30 minutes each;
- Speakers switched roles, speaking their native language in one conversation, English in the other;
- Conversations were lightly constrained to predefined topics. Topics were used more to prime conversations than to act as constraints, and included topics such as sports, pets, family, education, food, etc.

We recorded over 100 speakers for each language, with 50+ pairings. Speakers were balanced for gender and age groups. The English side of the recordings for Japanese and Chinese bilingual conversations were discarded as they represented accented speech, and thus less desirable, given our *unrestricted* requirement. In other words, the bilinguals we recruited were dominant in Japanese and Chinese first, English second. For English-only, we collected data from monolingual English conversations between speakers of different English dialects (American, Australian, British and Indian), ensuring speaker and dialect diversity.

2.2 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]`.

[2] For capturing the audio, we used a specially modified Skype client that allowed us to record the audio on the local computer, at the same time that they were holding a conversation.

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/>"`.[3]
- **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

The Microsoft Speech Language Translation (MSLT) corpus for Japanese and Chinese will be made available at the following URL:

- `https://aka.ms/mslt-corpus`

[3] In the latter example, taken from real data, the `<LM/>` tag indicates an utterance in a language that the transcriber did not know, and was left untranscribed, *e.g.*, I visited Ceuta.

Language	Data set	Files	Runtime	Average
English	Test	3,304	4h03m58s	4.4s
	Dev	3,052	3h56m37s	4.7s
Japanese	Test	4,160	5h22m23s	4.6s
	Dev	3,179	4h29m07s	5.1s
Chinese	Test	1,285	2h24m36s	6.8s
	Dev	1,256	2h12m28s	6.3s

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

At this site, we provide the audio files (see format description in Section 3.1 below), disfluent and fluent transcripts ("T1" and "T2"), and English translations ("T3"). In addition to the English, Chinese and Japanese corpora, we provide links at this site for the other corpora in the MSLT family of corpora, including the English, German, and French MSLT corpus released last year [17]. We ask that users of the Japanese and Chinese corpora cite this paper when used in their research (and [17] when using the English, French and German corpora). Also, please refer to the license agreement contained in the download packages for details on citation and limits of use.

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.

3.2 Transcription and Translation 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:** represents 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. Audio were provided to annotators segmented at the utterance level. Segmentation was done using an existing ASR engine using a Voice Activity Detection (VAD) algorithm. We observed bias when speakers annotated their own transcripts (repairing, *e.g.*, disfluencies and restarts, or transcribing words based on original intent), so we assigned work to a different set of consultants to prevent this issue. The extra effort regarding transcription resulted in higher transcription fidelity, especially regarding disfluencies, noises and incomplete utterances. Both punctuation and case information are optional in T1 but we found that most annotators already provided this. We assume they added this information to make the subsequent T2: Transform processing easier.

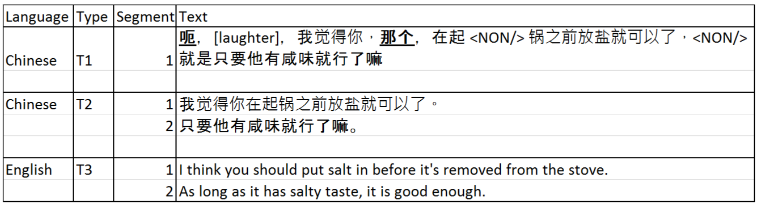
- **T2: Transform:** represents a cleaned up version of the T1 transcript with proper punctuation and case information. Of course, T2 data should not contain any disfluencies or other annotations. T2 output also should be segmented into *semantic units*. While the audio signal has already been segmented using VAD, the resulting utterances typically contain multiple phrases instead of a single sentence. This is partly due to the human speech production process and partly due to deficiencies in our speech segmentation. As machine translation targets individual input sentences, the T1-to-T2 segmentation process is crucial. 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. Annotators work on the T1 text files only and do not have access to the original audio files. We found that giving the annotators access to the audio signal resulted in longer annotation times, sometimes contradicting the original T1 data, and with less focus on the transformation task.
- **T3: Translate:** represents the translation of the fluent T2 transcript. The goal is to create conversational target text which feels natural to native speakers. Every translation should be usable in a direct quote in a newspaper article. 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. The biggest problem for translators was lack of context. Especially for shorter utterances, we observed a lot of ambiguity which made the translation process hard. While we sent out T2 data in order (so that translators could have used contextual cues), any kind of task parallelization will have negatively affected the translation process. Also, our assumption that unique source segments should always have the same target translation might not hold in the case of ambiguous, context-dependent phrases. Our lessons learnt during the original translation process will guide future translation campaigns creating additional references for this data set.

3.3 Corpus Statistics

Table 3 provides an overview on segment, token and type counts for both Test and Dev data for English, Japanese and Chinese. Token length for disfluent T1 transcripts and segmented, fluent T2 transcripts show expected behavior: segment counts increase and the token numbers decrease. 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 Japanese or Chinese.

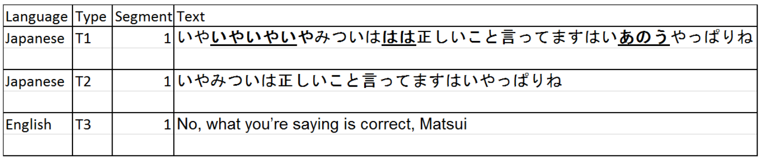
3.4 Some Examples

Figures 2 and 3 give examples containing disfluent, verbatim transcripts (T1), cleaned up and transformed text (T2) and the corresponding translations (T3) into English from both Chinese and Japanese. Note in the Chinese example how T2 transformation breaks the T1 transcript into two segments and also removes disfluencies and annotations. Translations are aligned on the segment level, and only with T2.



Language	Type	Segment	Text
Chinese	T1	1	**呃**，[laughter]，我觉得你，**那个**，在起 <NON/> 锅之前放盐就可以了，<NON/> 就是只要他有咸味就行了嘛
Chinese	T2	1	我觉得你在起锅之前放盐就可以了。
		2	只要他有咸味就行了嘛。
English	T3	1	I think you should put salt in before it's removed from the stove.
		2	As long as it has salty taste, it is good enough.

Figure 2: Examples from the Chinese corpus, with raw transcripts, transformed transcripts, and translations into English. Disfluencies that are removed are highlighted in the source.



Language	Type	Segment	Text
Japanese	T1	1	いや**いやいやい**やみついは**はは**正しいこと言ってますはい**あのう**やっぱりね
Japanese	T2	1	いやみついは正しいこと言ってますはいやっぱりね
English	T3	1	No, what you’re saying is correct, Matsui

Figure 3: Examples from the Japanese corpus, with raw transcripts, transformed transcripts, and translations into English. Disfluencies that are removed are highlighted in the source.

4 Usage Scenarios

We have previously described the three levels of annotation for the MSLT corpus data. In this section, we will describe how one could use the different annotation layers and explain why all three are needed to evaluate end-to-end quality of a speech translation system.

4.1 Using T1 data: “Bilingual” Speech Recognition

First, our data allows one to measure quality for *bilingual* speech recognition. While the recorded speech data itself is monolingual, our recording setup was bilingual by design. In any given session, both speakers were native speakers of the non-English language, so they could natively understand one other. However, as one of the two had to give answers in English, an additional bilingual element was added to the conversation flow. This affects the conversation. Most notably, we observe a decreased number of words uttered compared to purely monolingual conversations, which makes our data special in this regard and naturally representative of bilingual conversations (rather than monolingual conversations).

Testing speech recognition quality with our data will typically be implemented using word error rate (WER) scoring, comparing an ASR hypothesis against our reference transcription. Depending on the output style of the ASR engine under investigation, the reference text is either T1 or T2 data. Many ASR systems will remove disfluencies and partial recognitions to make resulting transcripts more readable to humans. If testing against such a system, our T2 data should be used as reference for WER scoring. As the segmentation of the T2 reference transcripts will likely not match that of the ASR output (which might not be segmented at all), ASR output should be compared to the "joint" T2 reference, which is the concatenation of all T2 segments into a single line.

Of course, if the ASR system being evaluated does no disfluency processing, then the T1 transcript should be used as the reference for calculating WER.

4.2 Using T2 data: Disfluency Removal

In the construction of the MSLT corpus, we have put in extra effort to annotate and transcribe disfluencies and other non-speech sounds, which are common in conversational speech. Such annotations can be used to evaluate the quality of disfluency removal (DR) models. While it is possible to train machine translation models to learn to translate or otherwise deal with such phenomena—this works pretty well for simple disfluencies, but becomes far more challenging for non-obvious disfluencies or partial utterances and restarts; see [9] for an example of such a system—the data space for these is very sparse. Therefore, we found it more practical to apply a DR component to "clean up" our ASR output before translation, as a normalization step [18].

For evaluation of such DR systems, one would feed the disfluent T1 transcripts into a disflueny removal system and compare the resulting output to the fluent T2 transcripts, which act as the reference. As we have previously discussed, T2 data is both cleaned up (with respect to disfluencies or non-speech noises) and segmented into units that contain at least some semantic value. Doing this will affect the usefulness of the T2 data as references for disfluency removal. If the DR model also performs segmentation, then its output can be directly compared to the T2 references. It has to be noted, however, that even small differences in segmentation will negatively affect the comparison. Hence, it might make more sense to compare the DR output and T2 segments on a non-segmented level. This is similar to the problem of testing "fluentized" ASR output against T2 references, as mentioned above.

4.3 Using T3 data: Conversational Translation

Considering evaluation of machine translation, the main difference to existing test data lies in the conversational nature of the collected data. We are not aware of any data sets which have been produced following the same "bilingual" recording setup. While there are test sets based on conversational speech transcripts, they are typically based on monolingual conversations. Hence, they might not be ideal for testing of bilingual (or even multilingual) conversational MT[4]. The MSLT data is different here as it puts

[4] There are a host of reasons why this might be true: directionality bias (given that one would be translating content from one language to another in one direction but not the other); unnatural

the focus on such bilingual conversation scenarios, albeit emulating a perfect translation component in the form of speakers understanding the non-English language natively. As this approach represents an upper bound on achievable translation quality (subject to the individual language competency of the speakers), the resulting references are perfectly suited for evaluation of conversational translation.

The MSLT corpus data can also be used to evaluate machine translation quality for conversational speech transcripts. To do this, one would use the fluent and segmented T2 transcripts as input segments for an MT system[5]. The resulting output data would then be compared to the corresponding T3 references, using automated metrics such as BLEU or human annotation. As directionality matters for MT evaluation, we provide test sets for translation from English as well as for translation into English. It is important to note that the transcripts for these are from different recording sessions which have been conducted by different speakers. As instructions and recording setup were identical for these, we think that the resulting data represents high quality test data for evaluation of conversational MT.

4.4 End-to-end Speech Translation Systems

Next to testing the performance of components corresponding to the individual annotation layers in the MSLT corpus, its data can also be used for end-to-end testing of speech translation. The setup is straightforward: The system records spontaneous utterances from one or more participants of a conversation. The audio signal is then sent to the speech recognition component which creates disfluent, verbatim transcripts. In a follow-up step, a disfluency removal component removes any disfluencies and separates the input transcript into one or more segments. These segments are fluent and each corresponds to a single "semantic unit", as discussed earlier. In a last step, the fluent segments are translated into the target language. Translation quality is computed based on automated metrics or evaluated using human annotators.

4.5 Multimodal Translation

The MSLT corpus data may also be helpful for multimodal translation. This research area has recently seen increasing interest (as demonstrated by shared tasks at WMT 2016 and 2017 [19]; also as a keynote by Mirella Lapata at ACL 2017) and aims to solve translation problems based on multimodal input. Effectively, our data offers three different input layers (the audio files and the T1/T2 transcripts), all of which are mapped to a single output layer, the T3 translations. It may be possible to build a translation system which uses both the audio signal and the corresponding transcript (likely in a joint, neural network approach) to generate translation output. Quality of such translations can be evaluated using our data set.

conversational structures, words, and phrases in the target language; no equivalent set of disfluencies one sees in T1 transcripts; etc.

[5] Again, we point to [9] for an example of where noisy, disfluent transcripts were used as input in a conversational MT system. In such a setting, the MT system itself would be doing much of the disfluency processing, rather than some separate DR module. The upside of such a technique is that the MT system *could* produce relevant disfluencies *in the target language*, given bilingual conversational text data with such disfluencies represented in the data for both languages.

4.6 Evaluation Campaigns using MSLT

MSLT evaluation data for German, French and English was used in the Machine Translation and Speech Recognition tracks at IWSLT 2016[6]. Although participants had access to significant amounts of parallel training data, e.g., from the WMT campaigns[7], they had very limited parallel data for training conversational MT systems. The out-of-the-box MT systems trained on WMT data generally did poorly on MSLT and TED lecture test data. However, adapting the base models using held-out TED data showed significant improvements on both the MSLT and TED test data sets. A notable example are the results from the Karlsruhe Institute of Technology (KIT) submission to IWSLT 2016 [20], where adaptation led to more than 1.5 BLEU score improvements on the MSLT corpus, even though, as the authors noted, the MSLT corpus did not exactly match the TED data used for adaptation (which is lecture-focused, and less conversational).

		English		**Japanese**		**Chinese**	
Annotation	**Description**	**Test**	**Dev**	**Test**	**Dev**	**Test**	**Dev**
`<SPN/>`	Speech noise	200	271	1,445	1,122	722	694
`<EU/>`	End unspoken	409	388	43	32	20	15
`<NON/>`	Non-speech noise	192	235	1,446	1,192	987	1,077
`<UNIN/>`	Unintelligible	306	125	92	110	103	76
`<LM/>`	Language mismatch	12	0	0	0	44	56
`<AS/>`	Audio spill	6	0	0	0	11	0
`<SU/>`	Start unspoken	37	54	10	5	5	1
`<UNSURE/>`	Annotator unsure	59	81	36	28	24	22
`<NPS/>`	Non-primary speaker	44	68	3	2	36	27
`<MP/>`	Mispronounced	3	4	12	20	0	0
`[laughter]`	Laughter	217	192	31	26	55	43
	Annotations	1,487	1,418	3,471	2,801	2,057	2,067
	Utterances	3,304	3,052	4,160	3,179	1,285	1,256
	Tokens	42,852	41,450	9,169	4,985	3,751	3,804
	Types	36,318	35,308	8,413	4,964	3,510	3,597

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

5 Conclusion

We presented a corpus of Chinese and Japanese for end-to-end evaluation of speech translation systems and/or component level evaluation. In the latter case, the test data consists of component level data: to test the ASR component, albeit not relevant to MT *per se*, the corpus has audio data and verbatim transcripts; to test disfluency removal and related processing against the raw transcripts—a necessary component if one wishes to process “raw” transcripts coming from, say, an off-the-shelf ASR engine— the corpus

[6] http://workshop2016.iwslt.org

[7] http://www.statmt.org/

Language	Type	Segments	Tokens	Types	Segments	Tokens	Types
English	T1 (EN)	3,304	42,852	36,318	3,052	41,450	35,308
	T2 (EN)	5,175	36,388	31,981	5,313	36,184	31,960
	T3 (JA)	5,175	37,324	33,862	5,313	36,409	32,913
	T3 (ZH)	5,175	39,776	35,614	5,313	40,159	35,824
Japanese	T1 (JA)	4,160	9,169	8,413	3,179	7,333	6,689
	T2 (JA)	5,976	6,221	6,205	4,970	4,985	4,964
	T3 (EN)	5,857	36,853	34,461	4,965	28,105	26,399
Chinese	T1 (ZH)	1,285	3,751	3,510	1,256	3,804	3,597
	T2 (ZH)	2,156	2,208	2,205	2,018	2,097	2,097
	T3 (EN)	2,156	15,665	13,920	2,018	14,284	12,628

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

has transcripts that have been cleaned up of disfluencies, pause words, discourse markers, restarts, hesitations, laughter, and any other content not relevant to translation; and to test conversational MT, the corpus has translated transcripts into English. We also provide English source with the same characteristics, translated into both Chinese and Japanese. This provides data that facilitates research in conversational MT both into and out of these two languages. 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 Chinese and Japanese conversations was thrown out due to non-English accents, and that all kept English sessions were recorded separately. We feel that the test and dev data that we are providing will be of great use to the community interested in developing conversational MT systems in and out of the Chinese and Japanese languages.

References

[1] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst, “Moses: Open Source Toolkit for Statistical Machine Translation,” in *Proceedings of ACL*, Prague, The Czech Republic, 2007.

[2] J. Devlin, R. Zbib, Z. Huang, T. Lamar, R. Schwartz, and J. Makhoul, “Fast and Robust Neural Network Joint Models for Statistical Machine Translation,” in *Proceedings of ACL*, Baltimore, Maryland, 2014, pp. 1370–1380.

[3] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, J. Klingner, A. Shah, M. Johnson, X. Liu, ukasz Kaiser, S. Gouws, Y. Kato, T. Kudo, H. Kazawa, K. Stevens, G. Kurian, N. Patil, W. Wang, C. Young, J. Smith, J. Riesa, A. Rudnick, O. Vinyals, G. Corrado, M. Hughes, and J. Dean, “Google’s Neural Machine Translation System: Bridging

the Gap between Human and Machine Translation," 2016. [Online]. Available: https://arxiv.org/abs/1609.08144

[4] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proceedings of ICLR*, 2015.

[5] F. Seide, G. Li, X. Chen, , and D. Yu, "Feature Engineering in Context-Dependent Deep Neural Networks for Conversational Speech Transcription," in *Proceedings of ACRU, IEEE*, 2011.

[6] W. D. Lewis, C. Federmann, and Y. Xin, "Applying Cross-Entropy Difference for Selecting Parallel Training Data from Publicly Available Sources for Conversational Machine Translation," in *Proceedings of the IWSLT 2015*, Danang, Vietnam, December 2015.

[7] P. Wang and H. T. Ng, "A Beam-Search Decoder for Normalization of Social Media Text with Application to Machine Translation," in *Proceedings of NAACL-HLT*, Atlanta, Georgia, June 2013, pp. 471–481.

[8] J. Godfrey and E. Holliman, "SWITCHBOARD: Telephone speech corpus for research and development," in *Proceedings of the IEEE Conference on Acoustics, Speech, and Signal Processing*, San Francisco, March 1992, pp. 517–520.

[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

[10] 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

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

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

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

[14] 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

[15] 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

[16] T. Takezawa, E. Sumita, F. Sugaya, H. Yamamoto, and S. Yamamoto, "Toward a Broad-coverage Bilingual Corpus for Speech Translation of Travel Conversation in the Real World," in *Proceedings of LREC 2002*, Las Palmas, Spain, 2002.

[17] C. Federmann and W. Lewis, “Microsoft Speech Language Translation (MSLT) Corpus: The IWSLT 2016 release for English, French and German,” in *Proceedings of the IWSLT 2016*, Seattle, Washington, December 2016.

[18] H. Hassan, L. Schwartz, D. Hakkani-Tur, and G. Tur, “Segmentation and disfluency removal for conversational speech translation,” in *In Proceedings of INTERSPEECH 2014*, 2014, pp. 318–322.

[19] L. Specia, S. Frank, K. Sima’an, and D. Elliott, “A shared task on multimodal machine translation and crosslingual image description,” in *Proceedings of the First Conference on Machine Translation.* Berlin, Germany: Association for Computational Linguistics, August 2016, pp. 543–553. [Online]. Available: http://www.aclweb.org/anthology/W/W16/W16-2346

[20] E. Cho, J. Niehues, T.-L. Ha, M. Sperber, M. Mediani, and A. Waibel, “Adaptation and Combination of NMT systems: The KIT Translation Systems for IWSLT 2016,” in *Proceedings of the IWSLT 2016*, Seattle, Washington, December 2016.