MuST-C: A multilingual corpus for end-to-end speech translation

Integrantes:
Juan Cobos
Christian Moreira
Edwin Narvaez

Contenido

- Contexto del problema
- Estado del arte actual
- Problema que ataca la investigación
- Metodología usada
- Resultados más interesantes
- Conclusiones

Contexto del problema

Procesamiento neuronal de extremo a extremo a ganado un interés cada vez mayor en problemas de procesamiento de lenguaje natural: **Reconocimiento automático de voz** (ASR) y **traducción automática** (MT)

Tienen muchas **ventajas**:

- Reducción de la ingeniería necesaria para entrenar módulos separados.
- Posibilidad de evitar errores acumulativos
- Mejor modelado del lenguaje humano

Así mismo presenta ciertas **desventajas:**

- Requieren gran potencia informática
- *Disponibilidad de **corpus.***

Contexto del problema

También se han aprovechado ampliamente los conjuntos de datos de ASR y MT en la traducción del lenguaje hablado (SLT).

- **Enfoque en cascada**
  - A pesar de que el enfoque en cascada obtiene buenos resultados, el avance en ASR y MT han motivado recientemente un cambio en la investigación de SLT hacia un framework end-to-end y sus ventajas potenciales.
- **Enfoque end-to-end**

Estado del arte actual

Los datos existentes para traducción del lenguaje hablado (SLT) de extremo a extremo deben tener la siguiente forma:

(señal de audio, texto traducido)

Grabación limpia de una oración completa pronunciada por un solo hablante.

Traducción al idioma de destino.

Estado del arte actual

Corpus	Languages	Hours of speech
(Niehues et al., 2018)	En → De	273
(Kocabiyikoglu et al., 2018)	En → Fr	236
(Tohyama et al., 2005)	En ↔ Jp	182
(Paulik and Waibel, 2009)	En → Es	111
	Es → En	105
(Post et al., 2013)	En → Es	38
(Stüker et al., 2012)	De → En	37
(Shimizu et al., 2014)	En ↔ Jp	22
(Federmann and Lewis, 2017)	En ↔ Jp/Zh	22
(Bendazzoli and Sandrelli, 2005)	En ↔ It/Es	18
	It ↔ Es	
(Bérard et al., 2016)	Fr → En	17
(Federmann and Lewis, 2016)	En ↔ Fr/De	8
(Woldeyohannis et al., 2017)	Am → En	7
(Godard et al., 2017)	Mboshi → Fr	4

IWSLT18

LibriSpeech aumentado

Situación actual:

Existen recursos libres disponibles

Pocos corpus

Lenguajes limitados

Tamaño pequeño (horas)

Problema que ataca la investigación

SLT de extremo a extremo ha ganado popularidad recientemente en sus dos tareas principales: reconocimiento automático de voz (ASR) y traducción automática (MT).

La investigación en el campo tiene que enfrentarse a la escasez de corpus disponibles públicamente para entrenar redes neuronales ávidas de datos.

Mientras que las soluciones en cascada tradicionales pueden basarse en datos de entrenamiento de ASR y MT considerables para una variedad de idiomas, los corpus de SLT disponibles para el entrenamiento de un extremo a otro son pocos, pequeños y con una cobertura de idioma limitada.

Metodología usada - Creación de MuST-C

- **Origen de los datos**
  - TED TALKS
  - Tiempo máximo por orador: 18 minuto
  - Especificaciones tomadas en cuenta:
    - Velocidad máxima de lectura: 21 caracteres/segundo
    - Exposición con un máximo de 2 líneas por subtítulo
    - 42 caracteres para cada subtítulo
  - Muy importante la eliminación de ruido
  - Actualidad:
    - Más de 143000 traducciones
    - Más de 33000 voluntarios
    - 116 idiomas utilizados

Metodología usada - Creación de MuST-C

Pasos del método utilizado:

1) **Descarga de datos**
   - Se inició con 2693 charlas en 14 idiomas diferentes.
2) **Segmentación y alineación a nivel de texto**
   - División a nivel de oración de las transcripciones y las traducciones
3) **Alineación de audio a texto**
   - Alineación forzada en inglés del audio con su transcripción
4) **Filtración**
   - Alineación audio-texto para crear un YAML
5) **Extracción de funciones de audio**
   - Eliminar palabras utilizando la herramienta XNM

Resultados más interesantes

MuST-C es un corpus de traducción de habla multilingüe creado para abordar la escasez de recursos para entrenar enfoques de extremo a extremo hambrientos de datos para la traducción del lenguaje hablado.

MuST-C se construyó a partir de TED Talks en inglés, con el objetivo de combinar en un recurso único todas las características deseadas de un corpus SLT.

Resultados más interesantes

MuST-C es un corpus que cuenta con las siguientes características:

- Obtenido en base a una gran variedad de temas y oradores,
- Amplia cobertura de idiomas
- Gran tamaño
- Alto calidad
- Distribución gratuita.

Resultados más interesantes

Como se puede ver en la tabla, los modelos entrenados con datos de MuST-C logran mejores resultados en el conjunto de pruebas balanceadas en las tres tareas. Considerando estas mejoras de rendimiento como evidencia de la confiabilidad de la metodología de creación del corpus.

Table 5
Empirical verification results. ASR, MT and SLT systems are trained with English-German data produced with the IWSLT18 and the MuST-C corpus creation pipelines. Evaluation is performed on a mixed test set, which comprises data from the two pipelines in the same proportion. The evaluation metrics are WER for ASR and BLEU for MT and SLT.

Corpus	ASR (↓)	MT (↑)	SLT (↑)
IWSLT18	42.15	24.90	8.94
MuST-C	32.05	25.46	12.25

Resultados más interesantes

En particular:

- Una reducción de 10.1 puntos WER en **ASR** indica una mayor calidad de las alineaciones de transcripción de audio de MuST-C.
- Un aumento de BLEU de 0.56 puntos en **MT** indica una calidad ligeramente mejor para las alineaciones de transcripción-traducción
- Un aumento de BLEU de 3,31 puntos en **SLT** indica una mayor calidad de las alineaciones de audio translación.

Table 5
Empirical verification results. ASR, MT and SLT systems are trained with English-German data produced with the IWSLT18 and the MuST-C corpus creation pipelines. Evaluation is performed on a mixed test set, which comprises data from the two pipelines in the same proportion. The evaluation metrics are WER for ASR and BLEU for MT and SLT.

Corpus	ASR (↓)	MT (↑)	SLT (↑)
IWSLT18	42.15	24.90	8.94
MuST-C	32.05	25.46	12.25

Resultados más interesantes

Table 7
Examples of partial alignments. Audio-transcription pairs: description of the audio segment with respect to the corresponding transcription. Trancription–translation pairs: in bold the words missing in the other sentence.

1.	TRANSCRIPTION	***At*** *most, they'd stick a feather on somebody's nose, and if it twitched, they didn't bury them yet.*
	AUDIO	The word "At" is not included in the audio segment.
2.	TRANSCRIPTION	*I'm here today to talk to you about circles and* ***epiphanies****.*
	AUDIO	The word *"epiphanies"* is not included in the audio segment.
3.	TRANSCRIPTION	*The next step we took was to find the applications* ***for it****.*
	AUDIO	The word *"submitted"* (from the previous sentence) is incorrectly included in the audio. The words *"for it"* are missing in the audio.
4.	TRANSCRIPTION	*Here are two different jewelry displays.*
	TRANSLATION	Ecco due diversi gioielli. **Uno si chiama "Jazz" e l'altro "Swing"**. [Eng: *Here are two different jewelry displays. One is called "Jazz" and the other one "Swing"*]
5.	TRANSCRIPTION	***The icebergs around me were almost 200 feet out of the water,*** *and I could only help but wonder that this was one snowflake on top of another snowflake.*
	TRANSLATION	E non riuscivo a non meravigliarmi del fatto che fossero fiocchi di neve posatisi uno sopra l'altro. [Eng: *And I could only help but wonder that this was one snowflake on top of another snowflake.*]
6.	TRANSCRIPTION	*I'm here today to talk to you about circles and epiphanies.*
	TRANSLATION	Sono qui oggi per parlarvi di cerchi e di epifanie, **cioè manifestazioni**. [Eng: *I'm here today to talk to you about circles and epiphanies, that is manifestations.*]

Resultados más interesantes

El alto porcentaje de alineaciones correctas indica la buena calidad general de los datos de MuST-C para todos los pares de idiomas. En promedio en todos los pares de idiomas, las alineaciones correctas son respectivamente **90,85%** y **95,38%** para la transcripción de audio y la transcripción-traducción.

Table 6
Human evaluation results. Percentage of correct *audio-transcription*, *transcription-translation*, *audio-translation* alignments for each language direction.

	% of correct alignments		
	Audio ↔ Transcription (ASR dataset)	Transcription ↔ Translation (MT dataset)	Audio ↔ Translation (SLT dataset)
En-Ar	90.4	89.4	81.8
En-Cs	89.8	95.6	85.4
En-De	92.4	97.4	90.2
En-Es	91.6	96.2	88.0
En-Fr	91.8	99.6	91.4
En-It	92.2	98.2	91.2
En-Nl	90.6	97.6	88.2
En-Pt	91.4	97.2	89.0
En-Ro	92.6	98.0	90.8
En-Ru	91.8	95.6	88.0
En-Zh	85.0	84.4	73.8

Conclusiones

Este es un corpus cuyas principales características apuntan a sentar las bases para futuras investigaciones sobre SLT de extremo a extremo. La amplia cobertura lingüística y la diversidad que ofrece MuST-C ampliarán la posibilidad de investigación y aplicaciones de SLT de extremo-extremo para los idiomas que quedaron fuera de la investigación anterior.

MuST-C combina en un solo recurso todas las características deseadas de un corpus SLT. Sin embargo, su utilidad potencial depende de la calidad de la traducción de audio.

Conclusiones

Hasta la fecha MuST-C es el recurso más grande disponible públicamente de su tipo. En su versión actual, comprende la transcripción en inglés y las traducciones a 14 idiomas de destino de al menos 237 horas de habla por idioma.

MuST-C es un recurso en constante expansión y sus creadores afirman que se incluirán nuevos idiomas, mientras que se agregarán nuevos datos para los idiomas ya cubiertos.

Gracias por su atención.