Spoken Language Translation

Jan Niehues
Matteo Negri
Matthias Sperber
Sebastian Stüker
Marco Turchi

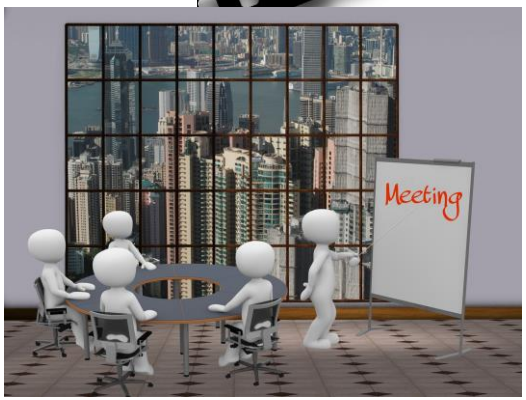
17/09/2019

jan.niehues@maastrichtuniversity.nl

Maastricht University

Use cases

- Presentations
  - Conferences/Lectures
- Videos
  - Internet: Youtube, Facebook, …
  - Television
- Every-day interactions
  - Tourist encounters, Medical care, Interactions with authorities
  - Telefon conversations
- Meetings

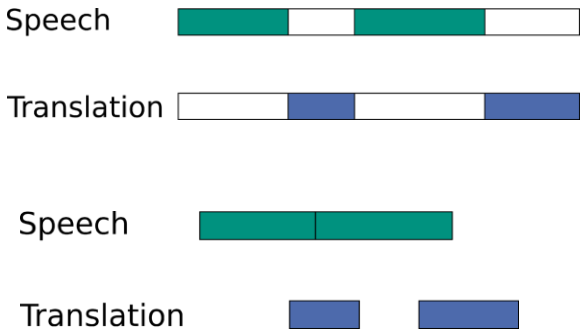


Overview

- Introduction
- Cascaded approach
- End-to-End Speech Translation
- Challenges:
  - Segmentation
  - Simultaneous translation
  - Spontaneous speech

Different Application scenarios

- Sequence
  - Consecutive translation
  - Simultaneous translation
  - Differences:
    - Segmentation
    - Speech overlap



Different Application scenarios

- Sequence
- Number of speakers
  - Examples:
    - Single speaker
      - E.g., presentations
    - Multiple speaker
      - E.g., meetings
  - Challenges:
    - Overlapping voice

Different Application scenarios

- Sequence
- Number of speakers
- Online/Offline systems
  - Offline: Translate audio in batch mode
    - E.g., movies
  - Online: Translate during production of speech
    - Real-time translations:
      - Translation as fast as speech input
    - Latency
      - Time that passes between speech and translation
      - Latency should be as minimal as possible

Different Application scenarios

- Sequence
- Number of speakers
- Online/Offline systems
- Presentation
  - Text
  - Audio
    - Additional TTS needed

Recent Data Resources

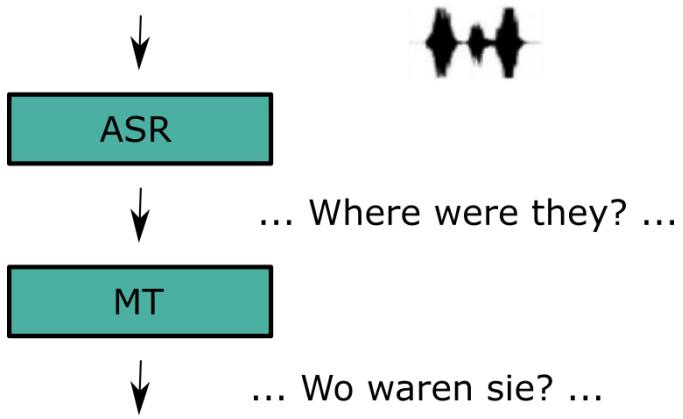
- Fisher data [Post et al., 2013]
  - Languages: Spanish to English
  - Domain: Telephone conversation
- MuST-C Corpus [Di Gangi et al., 2019]
  - Languages: English to 8 European Languages
  - Domain: TED
- LIBRI-TRANS [Kocabiyikoglu et al., 2018]
  - Languages: English to French
  - Domain: Audio books
- MASS [Boito et al, 2019], STC [Shimizu et al., 2014], BSTC, ..

Overview

- Motivation and Introduction
- Cascaded approach
- End-to-End Speech Translation
- Challenges:
  - Segmentation
  - Simultaneous translation
  - Spontaneous speech

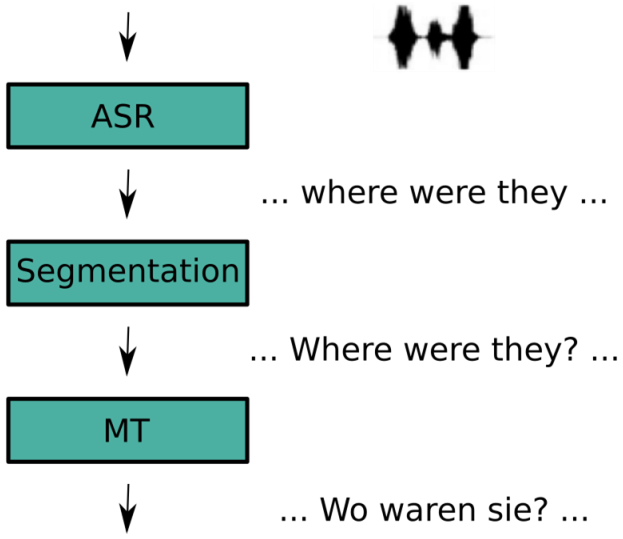
Cascade Spoken Language Translation

- Serial combination of several models
  - Automatic speech recognition (ASR)
  - Machine translation (MT)



Cascade Spoken Language Translation

- Serial combination of several models
  - Automatic speech recognition (ASR)
  - Machine translation (MT)
  - Segmentation
- Advantages:
  - Data availability
  - Modular system
  - Easy incorporation of new ASR/MT developments



Cascaded SLT: Challenges

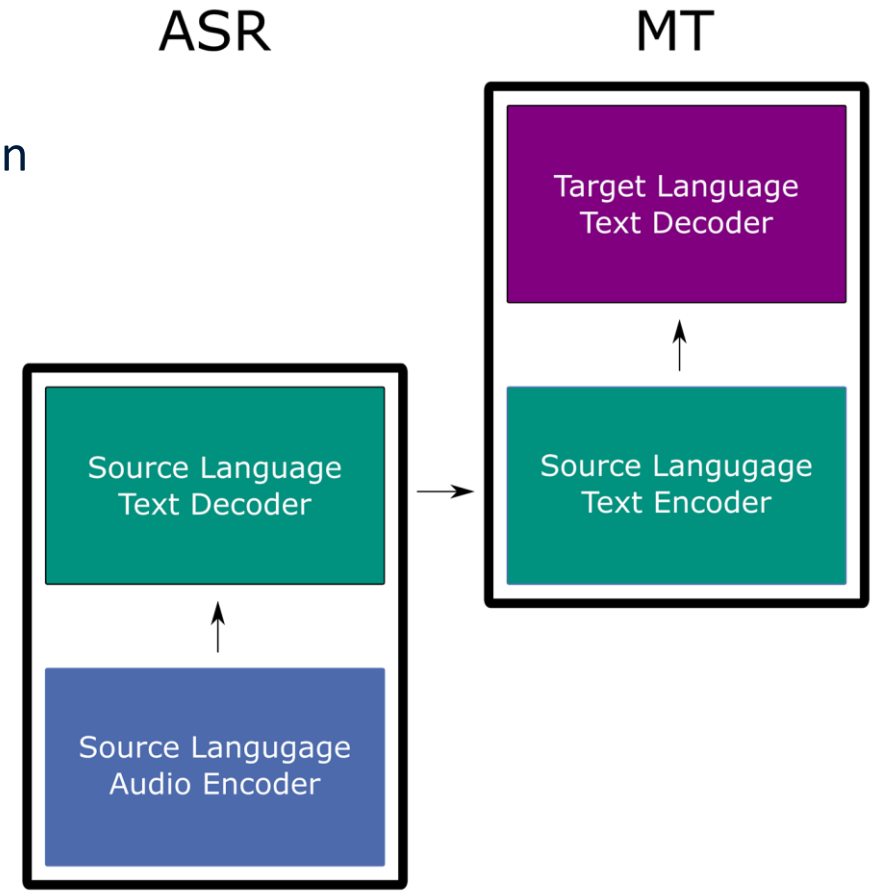
- Error propagation
  - Even the best components lead to errors
  - Solutions
    - Ignore
    - Represent different hypotheses
      - N-Best lists
      - Lattices [Saleem et al, 2005; Matusov et al, 2005]
    - Make MT robust to errors [Tsvetok et al. 2014; Lewis et al., 2015; Sperber et al, 2017]
- Separate optimization
- Script for source language is needed
- Computational complexity

Overview

- Motivation and Introduction
- Cascaded approach
- End-to-End Speech Translation
- Challenges:
  - Segmentation
  - Speech output
  - Simultaneous translation
  - Spontaneous speech

End-to-End SLT

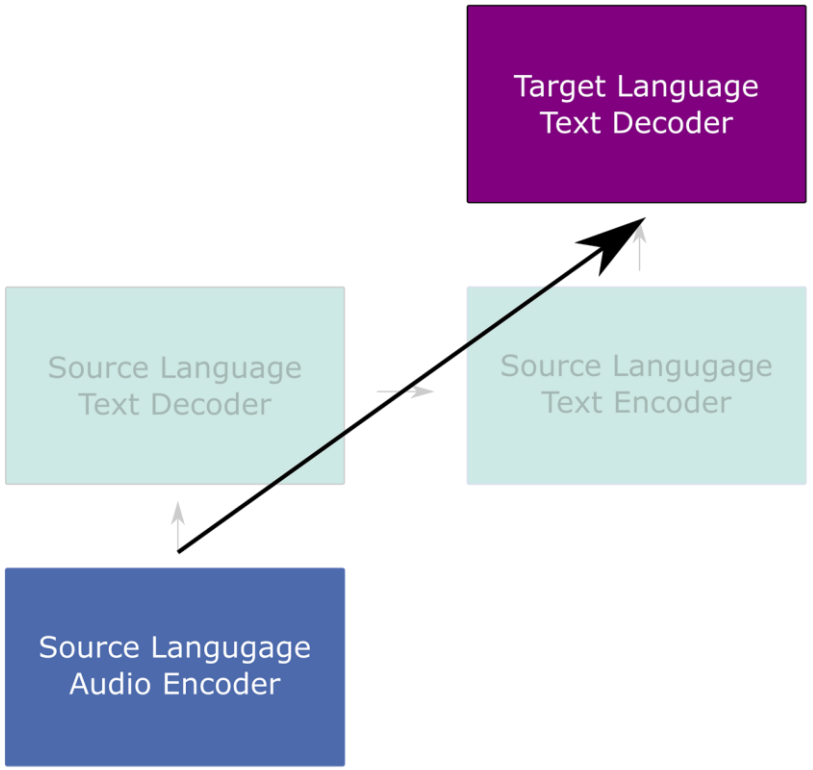
- Opportunity
  - Similar models for ASR and MT
    - Encoder/decoder with attention



End-to-End SLT

- Opportunity
- Directly learn mapping to target language text
  - [Duong et al., 2016;Berard et al., 2016; Weiss et al., 2017]

- IWSLT 2018 Evaluation:
  - Significant worse than cascaded models



E2E SLT - Challenges

- Input is audio signal
  - Longer sequences difficult to handle for NNs
  - Dependencies in time and frequency dimension
    - Approaches:
      - Apply techniques from automatic speech recognition
        - E.g. pyramidal encoder[Chan et al, 2016]

- Data availability
  - Few end-to-end speech translation corpora available
  - Often considerably smaller than MT and ASR training data
- Complicated mapping between source and target sequence
  - Source transcript can be intermedia supervised signal

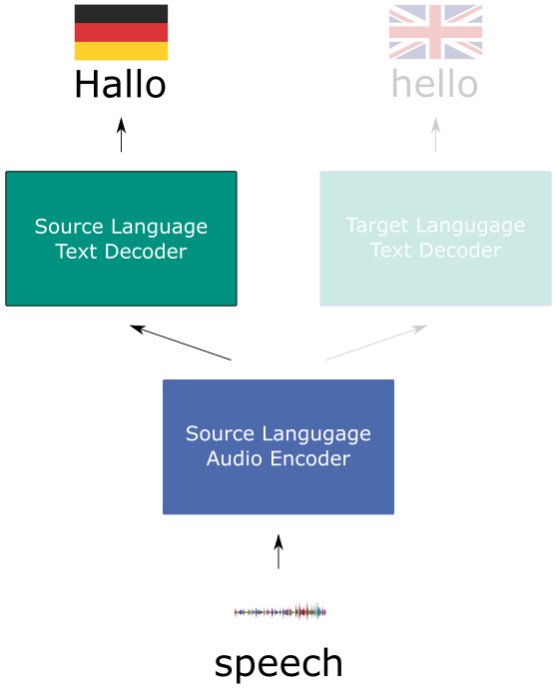
SLT Data

- Synthetic data:
  - Automatic generation by using TTS
    - [Berard et al, 2016; Kano et al, 2018;]
  - Challenge:
    - Generalization from TTS output to real audio signal

- Exploit other data sources by multi-tasking
  - Available data:
    - Speech data + transcripts
    - Parallel MT data
  - Idea:
    - Share parts of the network
    - Train SLT system using speech or MT data

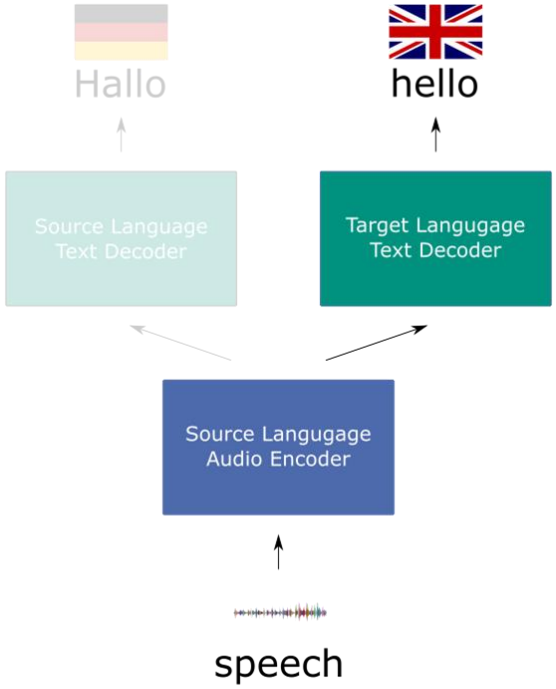
Multi-task learning

- Pre-training (Kano et al., 2018):
  - Train encoder on ASR task
  - Reuse on SLT task



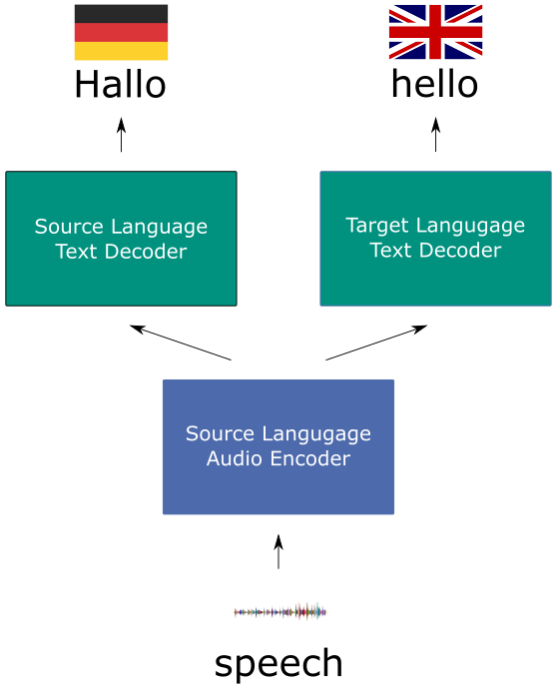
Multi-task learning

- Pre-training (Kano et al., 2018):
  - Train encoder on ASR task
  - Reuse on SLT task



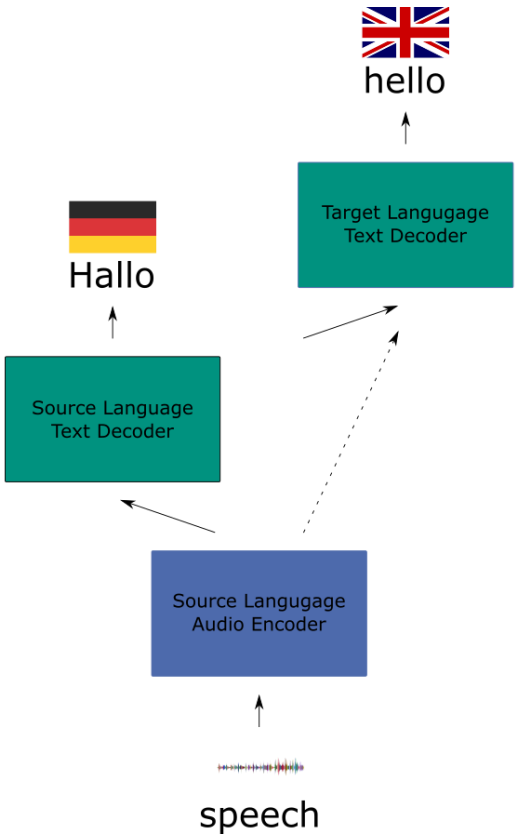
Multi-task learning

- Pre-training (Kano et al., 2018):
  - Train encoder on ASR task
  - Reuse on SLT task
- Multitasking (Weiss et al.,2017):
  - Train SLT and ASR jointly
- Challenge:
  - Data efficiency
  - How much gain from ASR/MT data?



2-stage NN Model

- SLT needs to learn complicated mapping
  - Supervised intermediate signal available
- Stack different decoders
  - Attend to source language decoder hidden states
- Triangle version:
  - Attend to source audio and source text [Anastasopoulos Chiang, 2018]
- Shared context vectors:
  - Ignore hard decisions of source language decoder [Sperber et al;2019]



Overview

- Motivation and Introduction
- Cascaded approach
- End-to-End Speech Translation
- Challenges:
  - Segmentation
  - Simultaneous translation
  - Spontaneous speech

Challenges - Segmentation

- Many applications:
  - Continuous audio stream
  - No punctuation in spoken language

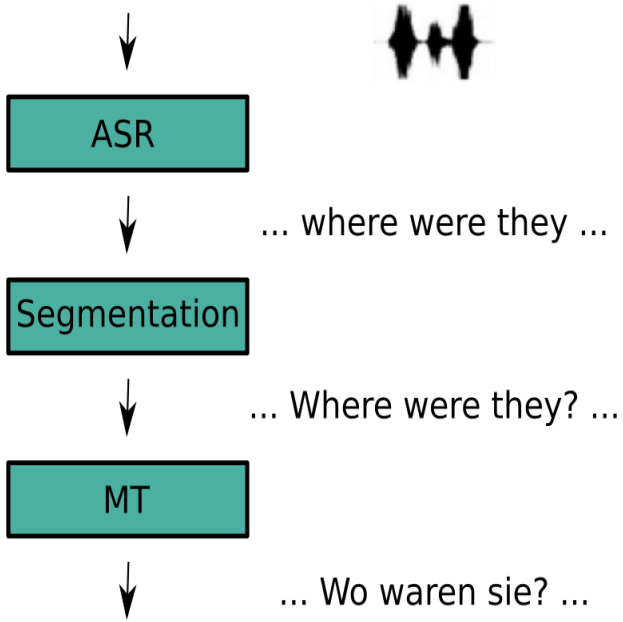
- Automatic segmentation and punctuation needed
  - Readability
  - Semantic
    - “Let’s eat Grandpa !”
    - “Let’s eat, Grandpa !”
  - Cascaded SLT:
    - MT often operates at sentence level

Challenges - Segmentation

- Add segmentation as additional component

- Approaches:
  - Language model-based
    [Stolcke et al, 1998; Rao et al, 2007]
  - Sequence labeling [Lu and Ng, 2010]
  - Monolingual machine translation
    [Peitz et al, 2011;Cho et al, 2012]

- Integration:
  - Between ASR and MT
  - After MT
  - Include into MT



Challenges – Simultaneous Translation

- Generate translation while speaker speaks
- Tradeoff:
  - More context improves speech recognition and machine translation
    - Wait as long as possible
  - Low latency is important for user experience
    - Generate translation as early as possible
- Challenge:
  - Different word order in the language
    - SOV vs SVO

German	Ich	melde	mich	zur	Interspeech	2019	an
Gloss	I	regester/ cancel	myself	to	Interspeech	2019	
English	I	????					

Challenges – Simultaneous Translation

- Approaches:
  - Learn optimal segmentation strategies
  - Stream decoding
    - Dynamically learn when to generate a translation
  - Re-translate
    - Update previous translation with better ones

Simultaneous Translation: Learn optimal segmentation strategies

- Idea:
  - Create segments that optimizing tradeoff between segment length and translation quality
- Advantages:
  - No changes to the NMT system
- Disadvantage:
  - Shorter context during translation
- E.g.:
  - Oda et al., 2014

Example:

Ich melde mich

zur Interspeech 2019 an

Simultaneous Translation: Stream decoding

- Idea:
  - At each time step:
    - Decided to output word
    - Wait for additional input

- Methods:
  - Dynamic decision (Cho et al, 2016; Gu et al, 2017; Dalvi et al, 2018)
  - Fixed schedule (Ma et al, 2019)

- Advantage:
  - Longer context into the past is available

- Disadvantage:
  - Major changes to the architecture
  - Balance between latency and quality

Simultaneous Translation: Re-translation

- Idea:
  - Directly output first hypothesis (low latency)
  - If more context is available
    - Update with better hypothesis (high quality)
  - Not only for MT, but for all components [Niehues et al, 2016]
  - Example:
    - Ich melde mich → I register
    - Ich melde mich von der Klausur ab → I withdraw form the exam
- Advantages:
  - Low latency and high quality
- Disadvantages:
  - Bad user experience if there are many updates
  - High computation cost

Challenges – Spontaneous speech

- Speech often spontaneous
  - Disfluencies

- Cascaded approach
  - Special model to generate clean text
  - E.g., as sequence labeling task [Cho et al, 2014]

- End to End:
  - Jointly learn to translate and remove speech disfluencies [Salesky et al, 2019]
  - Challenge:
    - Data resources

Summary

- Speech translation adds additional difficulties
  - Segmentation
  - Disfluencies
  - Simultaneous translations
- Cascade models often still state of the art
- Significant improvements in end-to-end models

Future research directions

- Simultaneous E2E Speech Translation
  - Segmentation
  - Stream decoding
- Different data conditions
  - Multilingual models
  - Low/Zero resource models
- Prosody
- Manual interaction

16th IWSLT 2019

Hong Kong

2nd - 3rd November 2019

16th International Workshop on
Spoken Language Translation

Important Dates:

Sep. 1: Paper Submission
July 1 - Sept. 8: Evaluation Period
Oct. 13: Acceptance - Notification

www.iwslt.org