Karlsruhe Institute of Technology

Open Source Toolkit for Speech to Text Translation

Thomas Zenkel, Matthias Sperber, Jan Niehues, Markus Müller, Ngoc-Quan Pham, Sebastian Stüker, Alex Waibel

Institute for Antrophomatics

Motivation

- Speech translation interesting challenge
  - Neural models
  - End-to-End models
- Provide a baseline
  - Cascade of several models
- Easy to extend
  - Develop models for part
- Easy to use
  - Download pretrained models

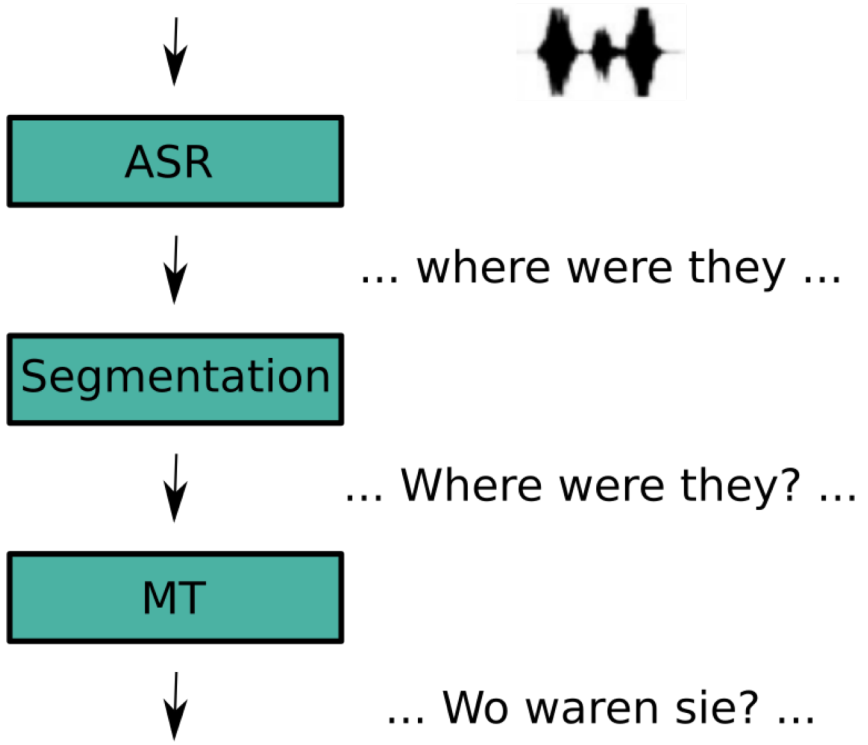
Cascade Spoken Language Translation

- Serial combination of several models

- ASR
  - Audio → Text

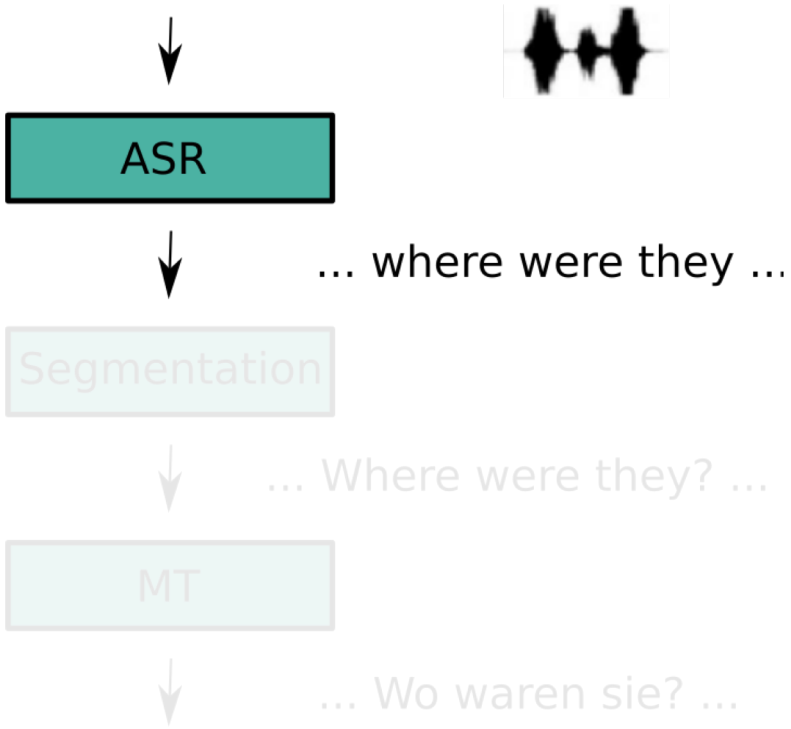
- Segmentation
  - Add case information
  - Add punctuation information

- Machine translation
  - Source language → target language



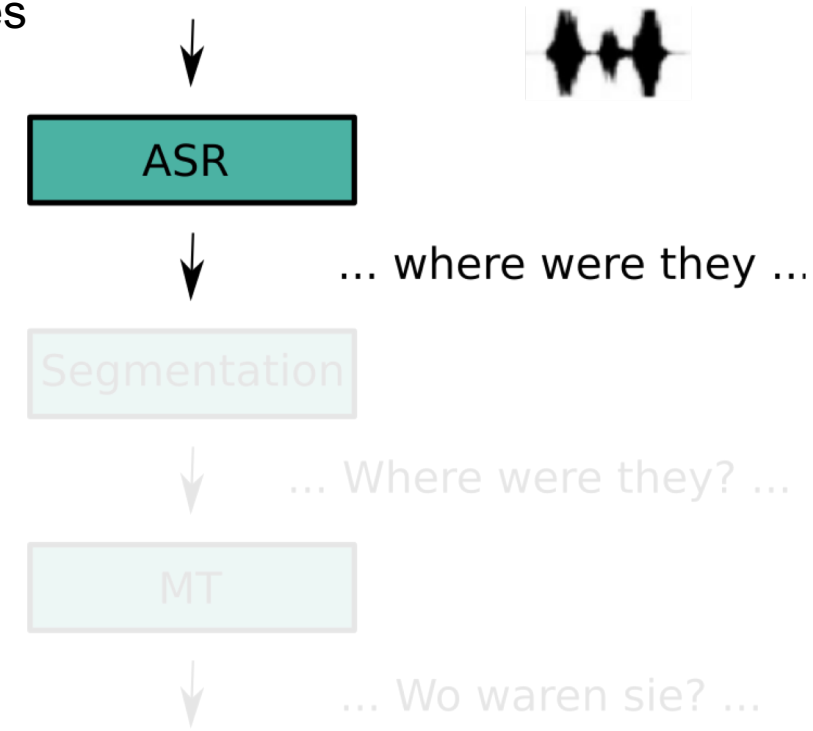
CTC-based ASR

- Input:
  - 40 dimensional Mel-filterbank features
- Output:
  - Byte-pair units (300 or 10000)
- Model:
  - 4-layer Bi-LSTM
  - Softmax layer
- Trained using CTC loss function



Encoder-Decoder Based ASR

- XNMT-based implementation
- Input:
  - 40 dimensional Mel-filterbank features
- Encoder:
  - 4-layer bidirectional pyramidal encoder
- Decoder:
  - One-layer bidirectional decoder



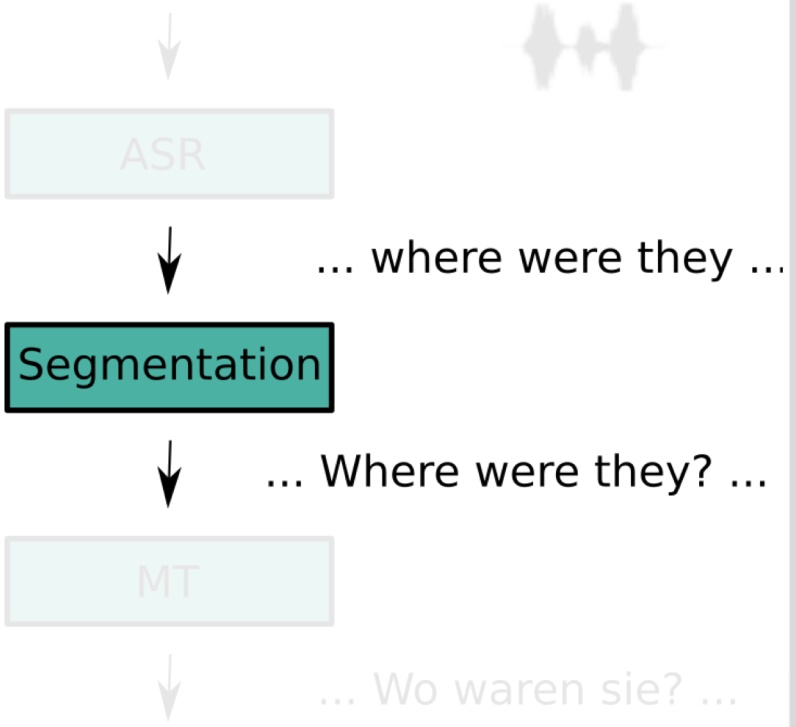
Segmentation and Punctuation

- Monolingual machine translation system
  - Add punctuation and case

- Example:
  - Input:
    - i felt wor@@ se why i wro@@ te a who@@ le book
  - Output:
    - U L L. U? U L L L L
    - I felt worse. Why? I wrote a whole book

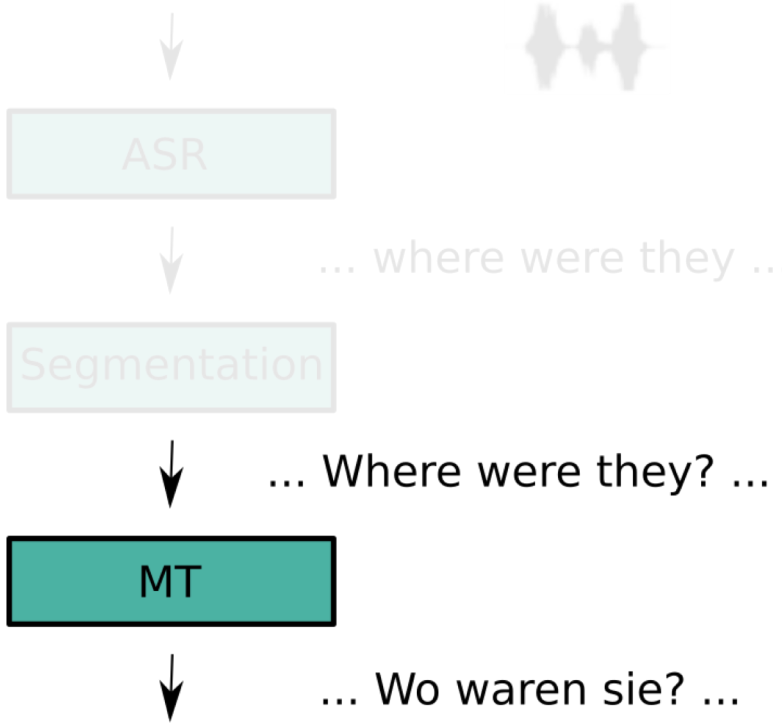
- Preprocessing:
  - Randomly split training data and remove punctuation information

- OpenNMT-based model



Machine Translation

- OpenNMT-based model
  - RNN-based Encoder and Decoder
- Preprocessing:
  - Tokenizer
  - Byte-pair encoding
- Mid-size model:
  - Pre-training on all data
  - Adaptation to in-domain data using continue training



Data

- Scripts to download and preprocess default data
- Audio:
  - TED LIUM corpus
- Text:
  - Small model:
    - WIT corpus
  - Midsize model:
    - EPPS corpus
    - WIT corpus

Results

- Evaluation tool to calculate 4 metrics provided
  - BLEU, TER, CharacTER, BEER
  - Automatic re-segmentation

Model	dev2010	tst2010	tst2013	tst2014
Attention	13.42	13.57	12.04	11.88
CTC 300	12.33	11.88	12.47	11.49
CTC 10K	13.04	13.44	13.41	12.58
Rover	13.98	14.08	13.73	13.23

Conclusion

- Combination of several toolkits to build full speech translation toolkit
- Easy usage:
  - Dockerized
- Applications
  - Apply pre-trained models
  - Train models using provided data (IWSLT)
  - Train models on own data
- Link:
  - https://github.com/isl-mt/SLT.KIT