Karlsruhe Institute of Technology

Speech Translation

Jan Niehues

Institute for Antrophomatics

Overview

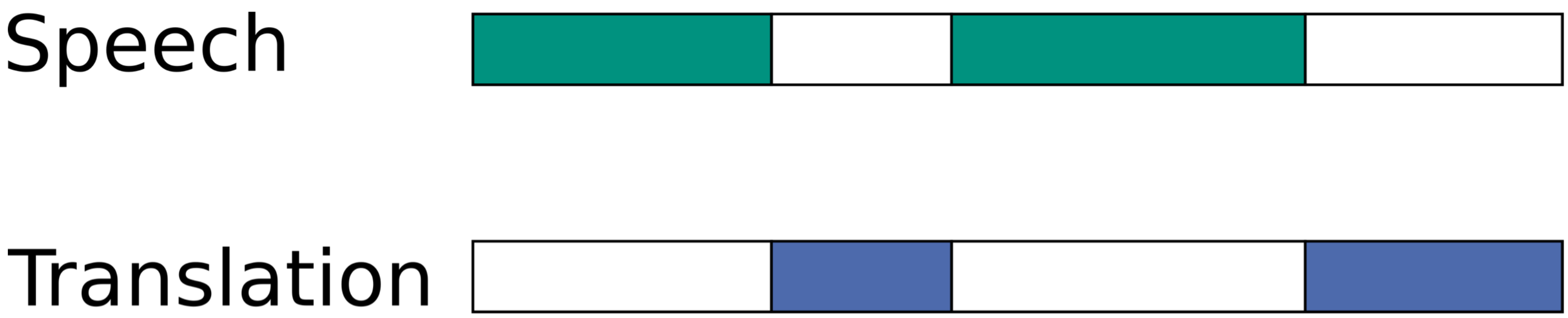
- Motivation
  - Challenges
- Cascaded approach
  - Automatic speech recognition
  - Machine Translation
  - Segmentation and Punctuation
- End-to-End Speech Translation
- Latency
- Disfluencies

Use cases

- Conferences / Lectures
- Internet videos
  - Youtube, Facebook, …
- Television
- Meetings
- Telephone conversations

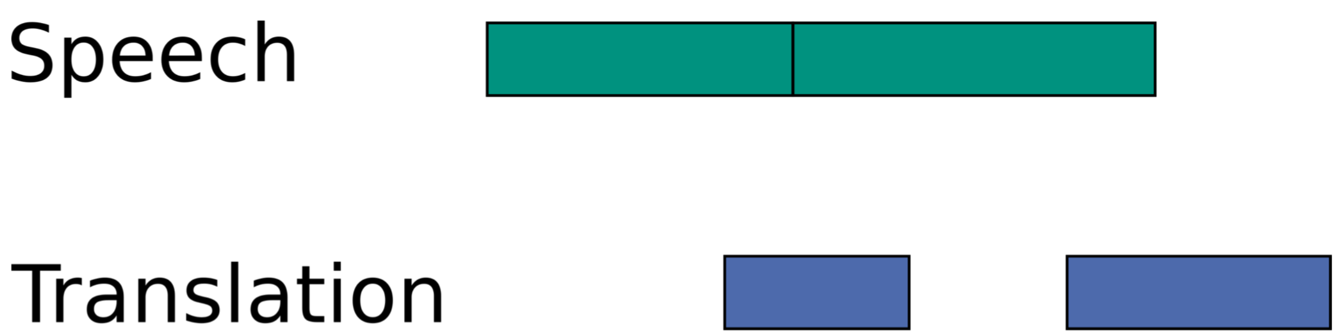
Different Application scenarios

- Sequence
  - Consecutive translation:
    - Speaker speaks a segment
    - Afterwards segment is translated
  - Characteristics:
    - Short Segments
    - Manual segmentation
    - Fixed dialog structure
      - No overlapping speech



Different Application scenarios

- Sequence
  - Simultaneous translation
    - Translation is provided while the speaker speaks
  - Characteristics:
    - Long segments
    - Automated segmentation needed
    - Flexible dialog structure



Different Application scenarios

Karlsruhe Institute of Technology

- Sequence
- Number of speakers
    - Single speaker
        - E.g. Presentations
    - Multiple speaker
        - E.g. Meetings
        - Challenges:
            - Overlapping voice
    - Mainly increases difficulty for speech recognition

Different Application scenarios

- Sequence
- Number of speakers
- Online/Offline systems
  - Online: Translate during production of speech
  - Offline: Translate full audio (e.g. movies)
  - Real-time translations:
    - Translation as fast as speech input
  - Latency
    - Time passes between speech and translation

History

- Speech translation systems for simple dialogs
  - Consecutive
  - Manual segmentation
  - Limited Domain

- Presentation translation
  - Simultaneous
  - Open Domain
  - Single speaker

- Meeting translation
  - Simultaneous
  - Multiple speaker

Challenges - Segmentation

- Segmentation:
  - No punctuation in spoken language
  - BUT punctuation marks are important

  - Let's eat Grandpa !
  - Let's eat, Grandpa !

  - Punctuation saves lives

Challenges - Segmentation

- Inserting correct punctuation marks difficult
  - Ambiguities

- Important hints:
  - Surrounding words
  - Context
  - Audio features
    - Pause
    - Pronunciation

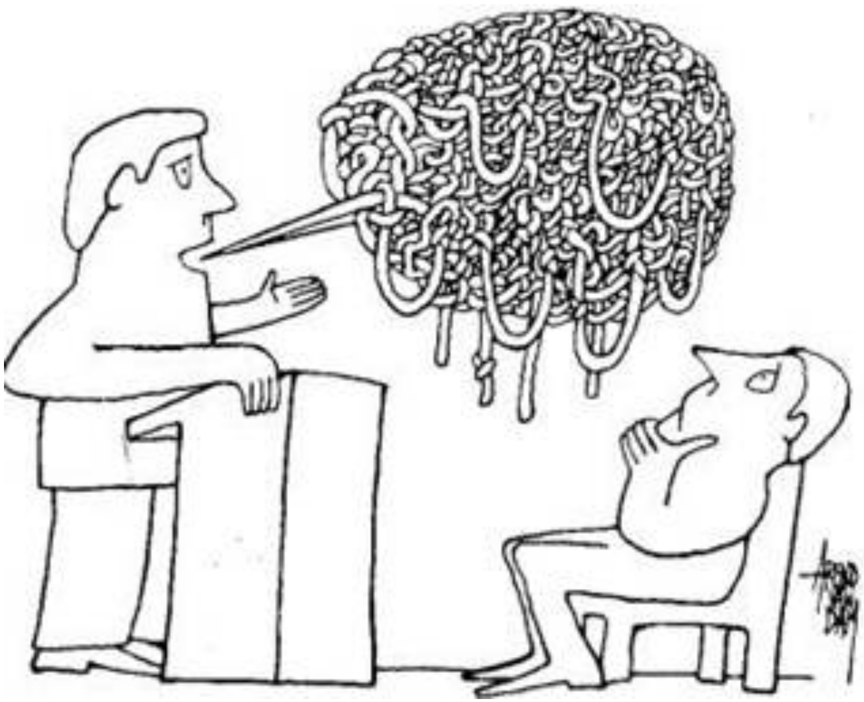
Challenges – Online Translation

- 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

- Approaches:
  - Automatically generate minimal segments
  - Dynamically learn when to generate a translation
  - Update previous translation with better once

Challenges – Spontaneous speech

- We are speaking spontaneously usually in our lives
  - Except for formal speeches, talk,…
- Almost all of speech in normal situations
- Speaker is not reading scripts
- Natural, relaxed
- Daily life
- Meetings, phone call
  - Multiple speakers



Characteristics of spontaneous speech

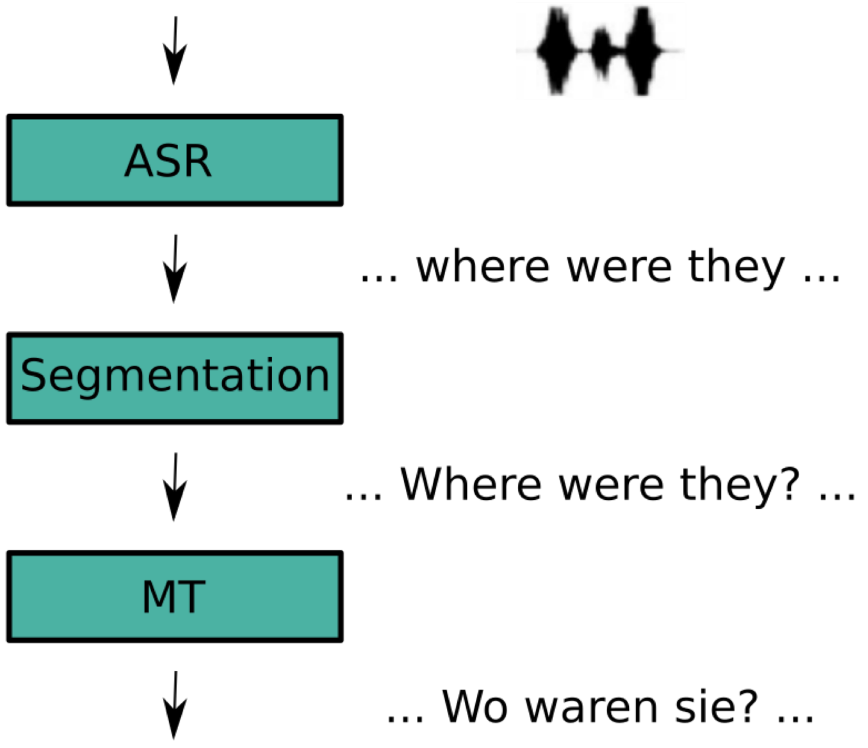
- Frequent use of filler words
  - “uh”, “uhm”, “hmm”
  - “ja”, “well”

- (rough) Repetition of phrases/words
  - “I mean, I mean I saw him there”
  - “there is, there was a cat”
  - “I would like to have a ticket to Denver, no, to Houston”

- Change of idea about what/how to speak
  - “We have here, uh, these fossils were discovered in Argentina…”
  - “How can you do that without, oh, what time is it now?”

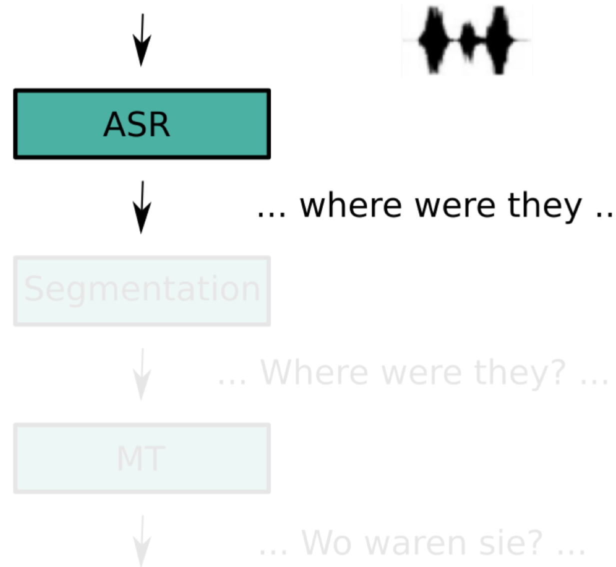
Cascade Spoken Language Translation

- Serial combination of several models
- ASR
  - Audio → Text
- Segmentation
  - Add case information
  - Add punctuation information
- Machine translation
  - Source language → target language



Automatic Speech Recognition

- HMM/DNN-based systems
  - Traditional ASR Systems
  - Still often state-of-the-art
- CTC-based Systems
  - LSTM to predict letters or blank symbol
  - CTC loss function
- Encoder-Decoder Systems



ASR Output

- Example:

  where
  were they and what did they
  talk about and now what was the topic of
  the discussion as this
  emotion of being angry came up now to be able
  to answer these questions you will
  also realize quite
  quickly that this of course…

  - Errors in segmentation
  - Often no punctuation
  - Often no case information

  - Difficult to read

ASR Output

- Segmentation and punctuation are improve for readability

  Where were they?
  And what did they talk about?
  And now what was the topic of the discussion, as this emotion of being angry came up?
  Now, to be able to answer all these questions, you will also realize quite quickly, that this of course…

How do segmentation and punctuation affect machine translation?

- **Translation output** of German to English translation system
- ASR

> > We see here is an example from the European Parliament, the European Parliament 20 languages
> > And you try simultaneously by help human translator translators the
> > Talk to each of the speaker in other languages to translate it is possible to build computers
> > The similar to provide translation services

- ASR + correct segmentation and punctuation added manually

> > We see here is an example from the European Parliament.
> > The European Parliament 20 languages are spoken, and you try by help human translator to translate simultaneously translators the speeches of the speaker in each case in other languages.
> > It is possible to build computers that are similar to provide translation services?

Segmentation and Punctuation

- Insertion of right punctuation gets difficult as the speech gets more disfluent

- Example:
  - “I (long pause) uh went to hair salon yesterday”

- Long pause can cause punctuation marks
  - “I.”
  - “uh went to hair salon yesterday.”

- For translation we need better segmentation and punctuation

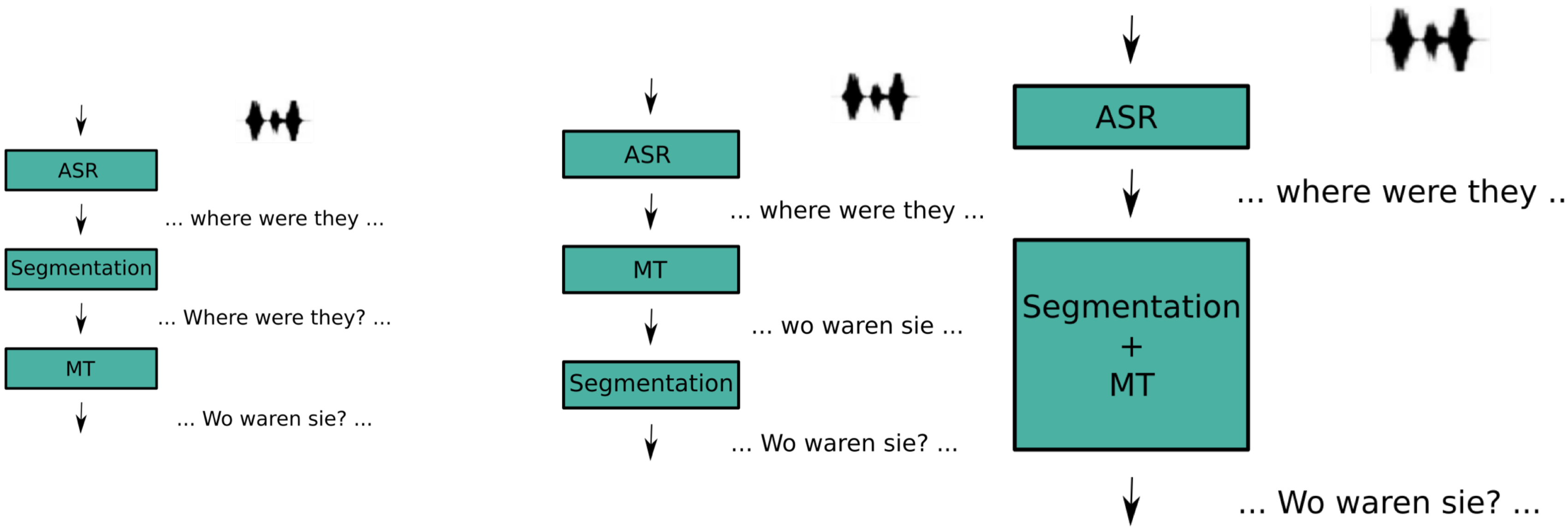
Affect of segmentation and punctuation in BLEU scores

Karlsruhe Institute of Technology

	BLEU
ASR	20.70
+ Segmentation	21.42
+ Full stop	22.18
+ All punctuations	22.48
Transcript	27.99

- For given German to English test set
- Segmentation and punctuation marks were added according to manual transcript
- All punctuations include: “?”, “!”, “,”,…

Adding Punctuation

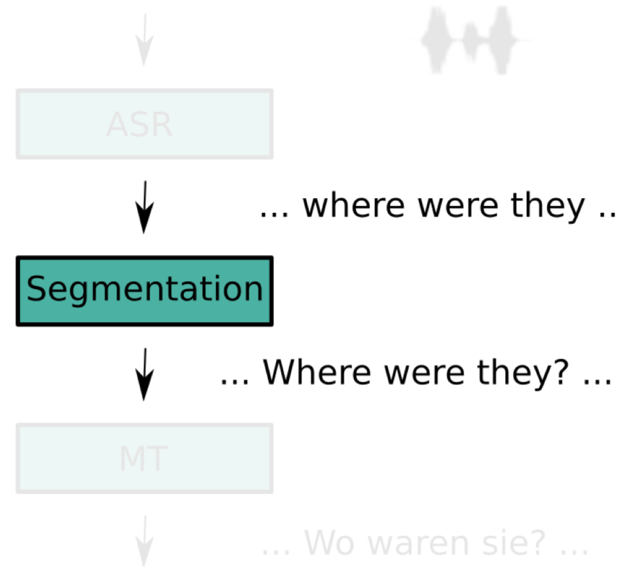


- Segmentation difficult in middle and right version

Segmentation

- Task:
  - Resegment text to sentence-like units
  - Insert punctuation marks
  - Often:
    - Correct casing of words

- Approaches:
  - Language model-based
  - Sequence labeling
  - Monolingual machine translation

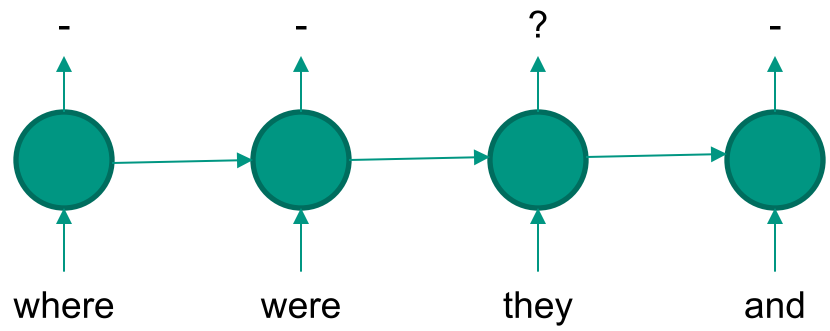


LM and prosody based model

- Consider two prior words and two after the possible punctuation marks
- LM trained on punctuated text
  - Score without an inserted punctuation mark
    - P(Hello Sir how are)
  - Score with a comma
    - P(Hello Sir , how are)
  - Score with a full stop
    - P(Hello Sir . how are)
- Pause longer than n seconds then a new segment
- Fast

Sequence labeling

- Input:
  - Sequence of words
- Output:
  - Following punctuation mark
- Models:
  - CRF, HMM, LSTM, …



Monolingual translation system

- Input:
  - Text without punctuation
- Output:
  - Text with punctuation
- Models:
  - Phrase-based SMT, NMT, …

- Steps:
  - Generate training data
  - Train model
  - Apply model to input data
  - Insert segment boundaries after punctuation

Monolingual MT- Training data

- Parallel text:
  - Remove punctuation from monolingual source text

Where were they
And what did they talk about
And now what was the topic of the discussion as this emotion of being angry came up
Now to be able to answer all these questions you will also realize quite quickly that this of course…

Where were they?
And what did they talk about?
And now what was the topic of the discussion, as this emotion of being angry came up?
Now, to be able to answer all these questions, you will also realize quite quickly, that this of course…

Monolingual MT- Training data

- Parallel text:
  - Remove punctuation from monolingual source text
  - Randomly split text

where
were they and what did they
talk about and now what was the topic of
the discussion as this
emotion of being angry came up now to be able
to answer these questions you will
also realize quite
quickly that this of course

where
were they? and what did they
talk about? and now, what was the topic of
the discussion, as this
emotion of being angry came up? now, to be able
to answer all these questions, you will
also realize quite
quickly, that this of course

Monolingual MT- Testing

- Sliding window to observe words in longer, various contexts

der	bildet	die	sogenannte	konjunktive	Normalform	wir	haben
bildet	die	sogenannte	konjunktive	Normalform	wir	haben	gesehen
die	sogenannte	konjunktive	Normalform	wir	haben	gesehen	dass
sogenannte	konjunktive	Normalform	wir	haben	gesehen	dass	wir
konjunktive	Normalform	wir	haben	gesehen	dass	wir	diese
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Monolingual MT- Testing

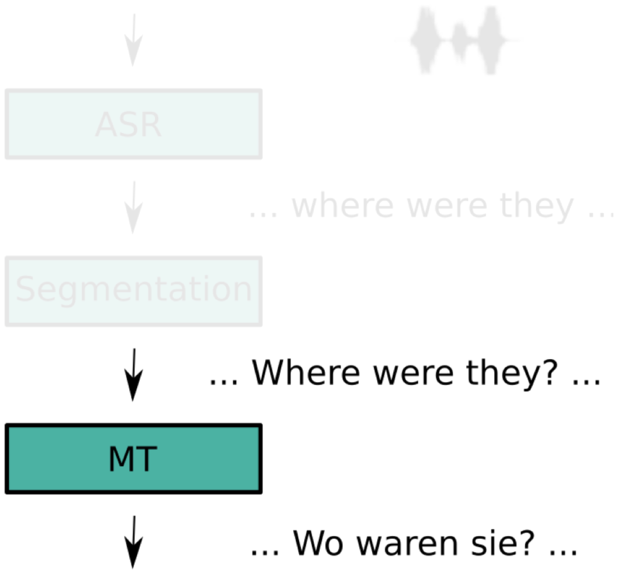
- Sliding window to observe words in longer, various contexts
  - Empirical threshold for inserting punctuation mark

der	bildet	die	sogenannte	konjunktive	**Normalform.**	Wir	haben
bildet	die	sogenannte	konjunktive	**Normalform.**	Wir	haben	**gesehen,**
die	sogenannte	konjunktive	**Normalform.**	Wir	haben	**gesehen,**	dass
sogenannte	konjunktive	**Normalform.**	Wir	haben	**gesehen,**	dass	wir
konjunktive	**Normalform.**	Wir	haben	**gesehen,**	dass	wir	diese
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Machine translation

- Baseline
  - Default NMT system

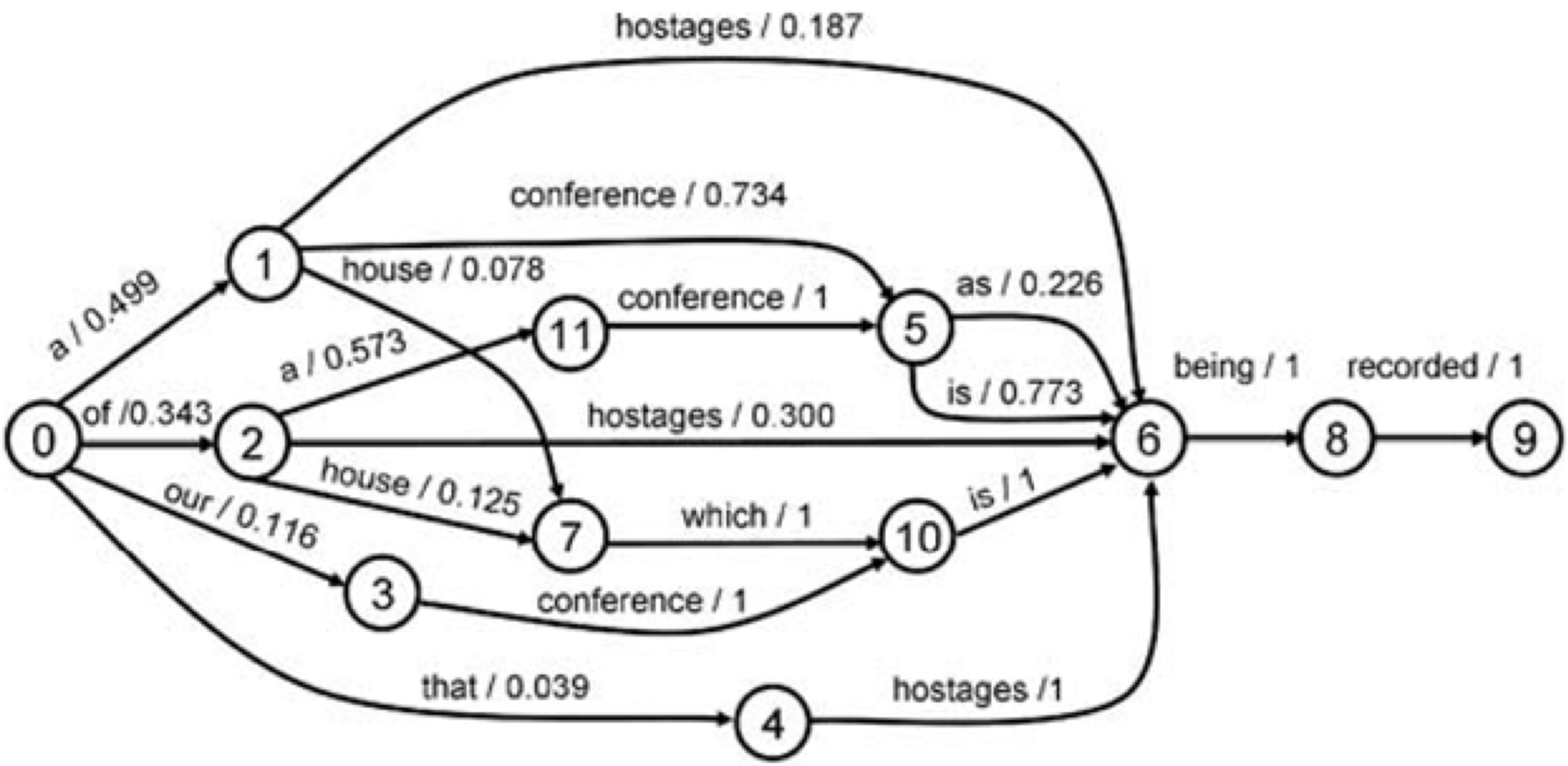
- Style in speech is different
  - Often adaptation to speech style
  - Continue training



ASR errors

- Even the best ASR system make errors
  - On difficult tasks even more
- MT has to deal with erroneous input
- Approaches:
  - Ignore
  - Tighter integration by using ASR lattices as input

ASR lattices



a conference is being recorded

Handling ASR Input

- Take first best ASR output
  - Problem:
    - No handling of ASR errors
  - Simple
  - Works often as good as other approaches
  - Reasons:
    - If the ASR system makes errors, it is hard for the MT to detect

Tight integration

- Use N-best output or ASR lattice as input for MT

…aber ausreichend für einfache Anwendungen **und des** Sie brauchten natürlich einen…

…aber ausreichend für einfache Anwendungen **und das** sie braucht natürlich einen…

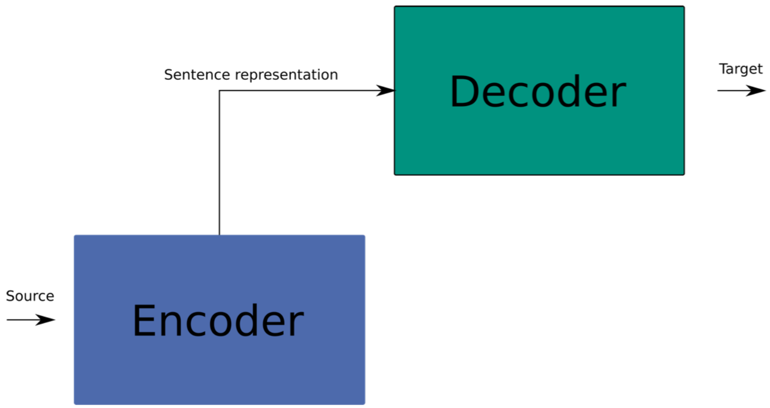
- Use score to model confidence of ASR system

- Problems:
  - MT might translate easier sentence, not correct one

End-to-End based systems

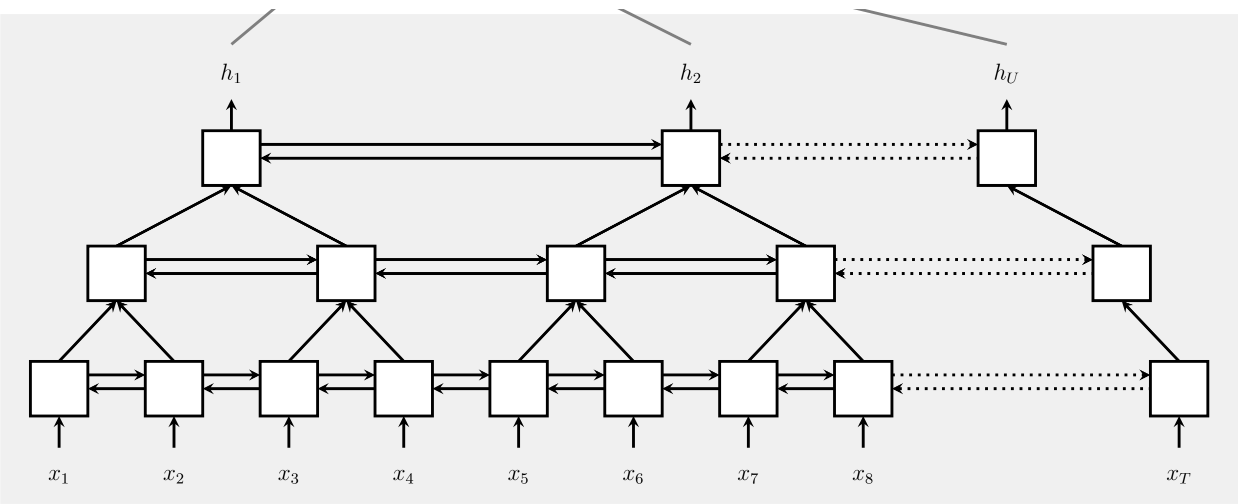
- Challenges of Cascaded systems:
  - Separated optimized Components
  - Hard to recover from ASR errors

- Opportunity:
  - Similar modelling of ASR and MT
  - Sequence to Sequence models



Attention-based ASR

- Main differences to Machine translation:
  - Encoder:
    - Larger input sequences
    - Reduce sequence length by Pyramidal encoder E.g.:
      - Concatenation/Summing of consecutive states
      - Convolution layer and stride at the bottom to downsample
    - Deep encoders



Chan et al. 2015

Attention-based ASR

- Main differences to Machine translation:
  - Encoder:
    - Larger input sequences
    - Reduce sequence length by Pyramidal encoder E.g.:
      - Concatenation/Summing of consecutive states
      - Convolution layer and stride at the bottom to downsample
    - Deep encoders
  - Decoder:
    - Character-based models

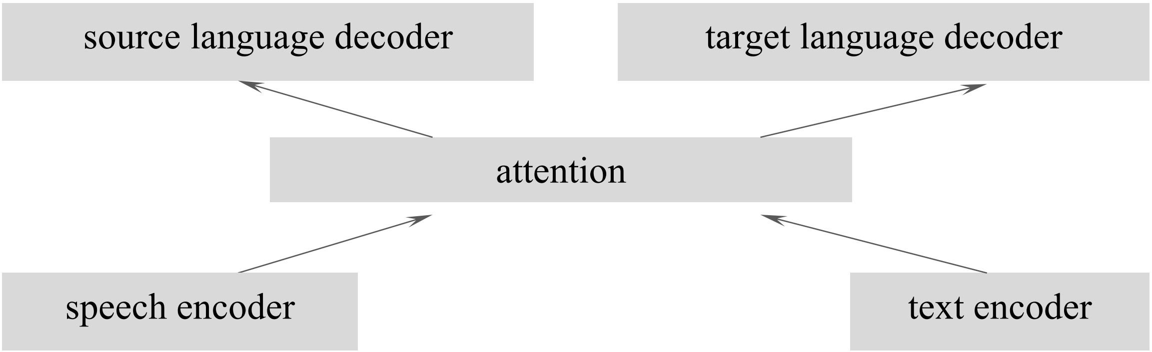
End-to-End SLT

- Encoder:
  - Source side audio encoder
- Decoder:
  - Character-based decoder with target language strings

- Results:
  - Mixed
    - Sometimes better/worse than cascaded

Challenges of End-to-End SLT

- Challenges of End-to-End SLT:
  - Rare direct end-to-end data available
- Idea:
  - Multi-task learning



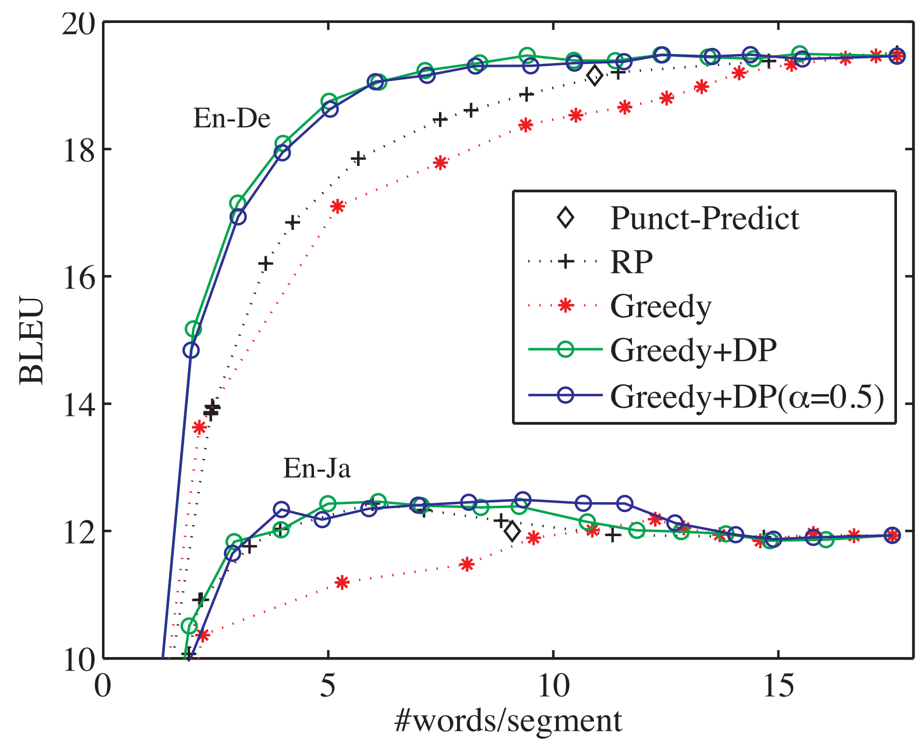
Latency

- Real-time spoken language translation
- The time between a word is spoken and when its transcript and translation are displayed to the user

- Each components adds to the latency
  - Computation time → fast servers with multiple cores, parallelized computations, smaller, faster models..
  - Communication time → fast connection, low overhead between components
  - Required context length?

Optimizing segmentation

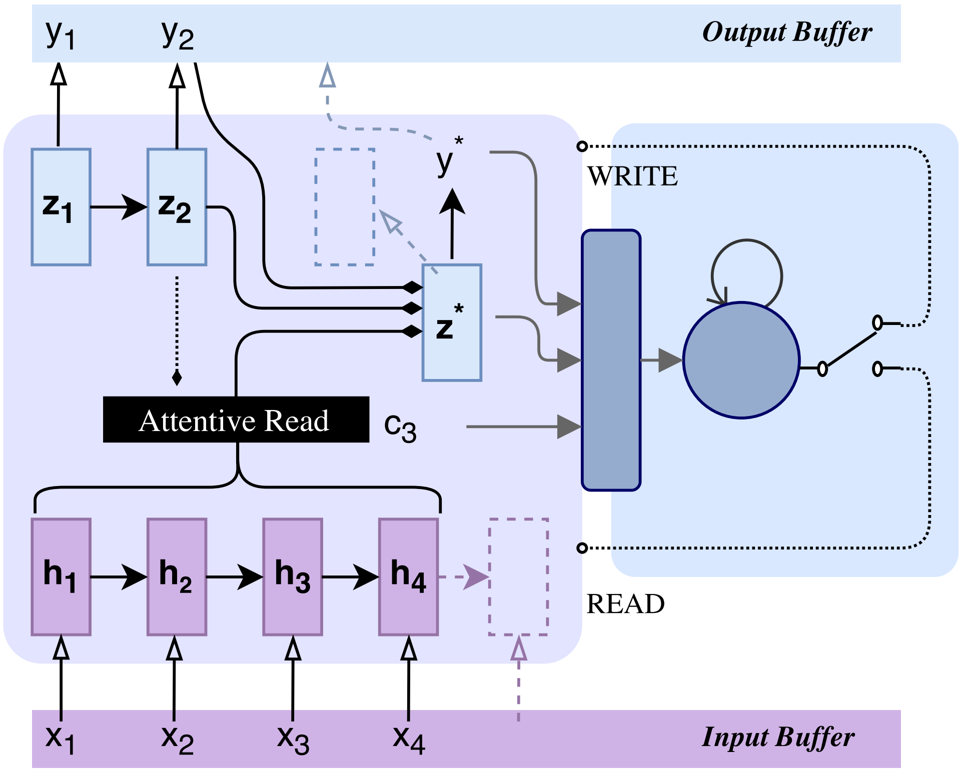
- Baseline:
  - Try to segmented into sentence
- Idea:
  - Create segments that optimizing tradeoff between segment length and translation quality



Oda et al., 2014

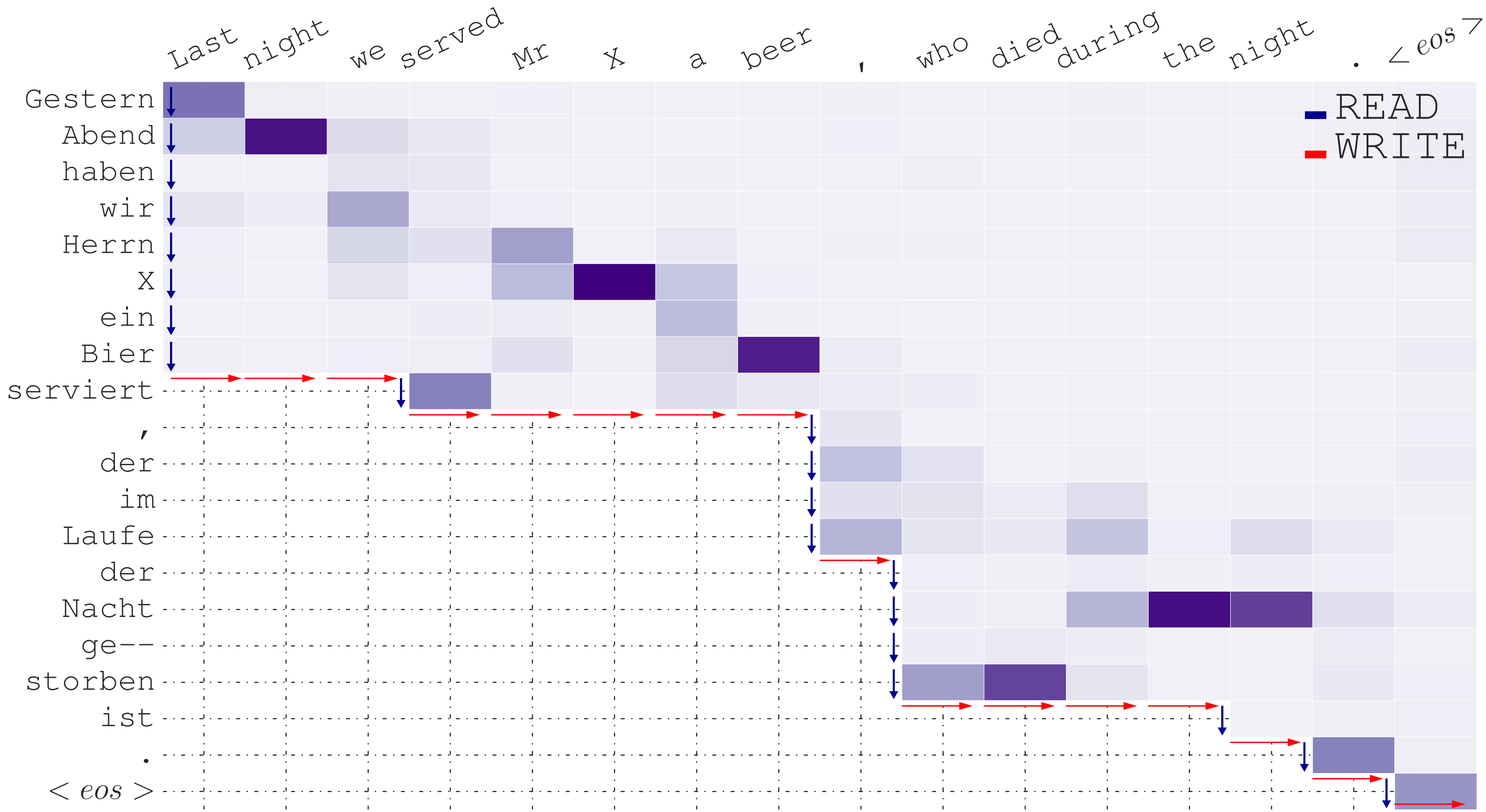
Jointly predicting Segments and Translation

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



Gu et al., 2017

Jointly predicting Segments and Translation



Gu et al., 2017

Updates of Hypothesis

- Directly output first hypothesis
- If more context is available:
  - Update with better hypothesis

- Example:
  - Ich melde mich
  - I register

  - Ich melde mich von der Klausur ab
  - I withdraw form the exam

- Not only for MT, but for all components

Updates of ASR

- Reduce the apparent latency
- ASR continually outputs its current best hypothesis e.g., once a second
- Updated by newer, possibly better, hypothesis
- Higher user acceptance than waiting for a complete, stable hypothesis

Example: Updates of ASR

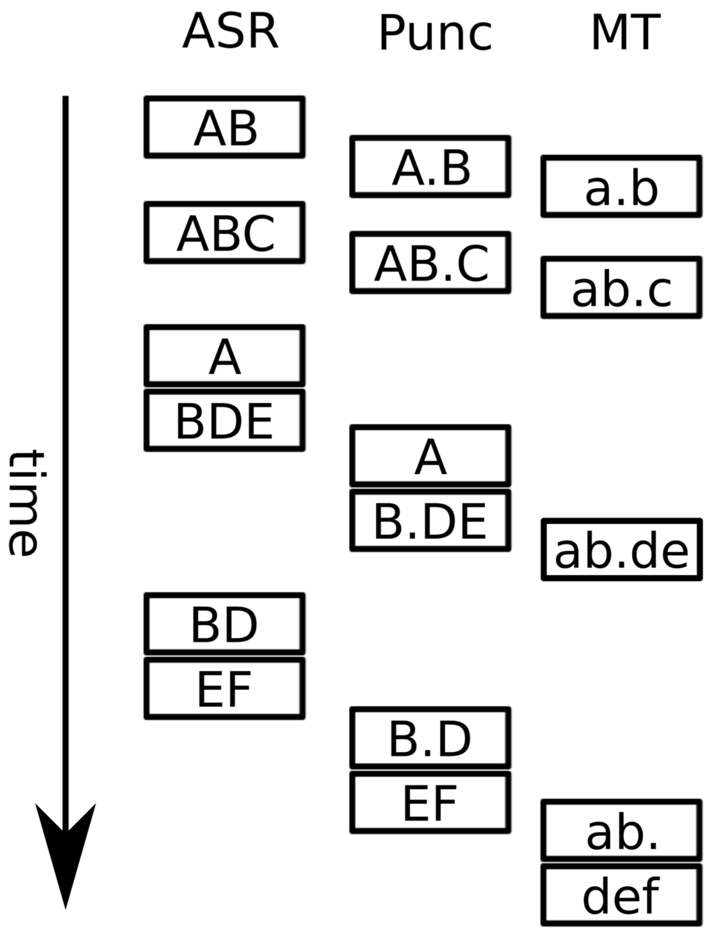
… In this planet you would have to prove ...

... In this planet you would have to provide 36 million translation ...

... Many dialects it is of course a dog ...

... Many dialects it is of course a daunting challenge ...

Update Protocol



- Difficulty:
  - Also input gets updated
- Message goes through the 3 components
- Hypothesis constantly getting updated

Results

- En→Fr
  - 7.5 average seconds → 1.8 seconds for initial output, 3.3 seconds for the final output
- De→En
  - 8.6 average seconds → 2 seconds for initial output, 5.3 seconds for the final output
  - Reordering
- Analysis

n	1	2	5	10	Full sentence	Update
Latency(s)	5.3	5.4	6.0	7.3	7.9	6.0
BLEU	8.5	9.3	10.2	11.2	11.4	11.4

  - Partial sentences (n words)
  - Same latency as n=5 system
  - Outperforms the same latency system by 1.2 BLEU

Challenges for NMT

- NMT will always generate full sentences

Input	Output
now,	ahora ,
now, I should	ahora debería , debería , debería .
now, I should men	ahora debería hombres hombres .
now, I should mention that this	ahora debería mencionar esto .

Challenges for NMT

- NMT will always generate full sentences
- Train also on partial sentences

Input	Output
now,	ahora ,
now, I should	ahora debería
now, I should men	ahora debería.
now, I should mention that this	ahora , debo mencionarlo .

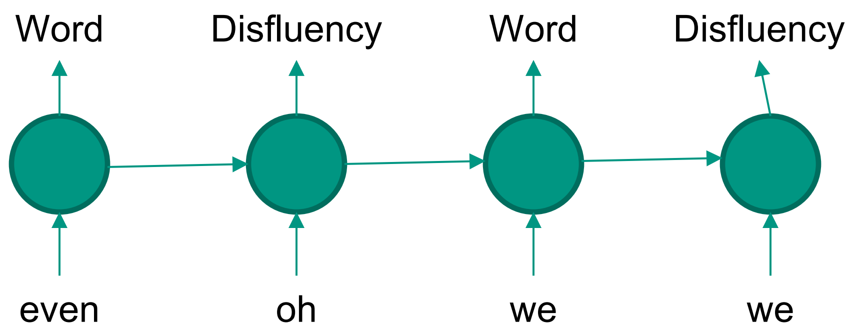
Disfluency

- Why is it so difficult?
  - Rough copies
    - The communication between man and machine, which we **customarily traditionally** always see, is the...
  - Some filler words, which can be filler, but sometimes not
    - “ja” in German
    - “well” in English
      - “we can’t even well we’re not even…”
      - “You did it very well”
  - Nearly no training data
  - ASR output may contain errors
  - Dangerous to remove to much

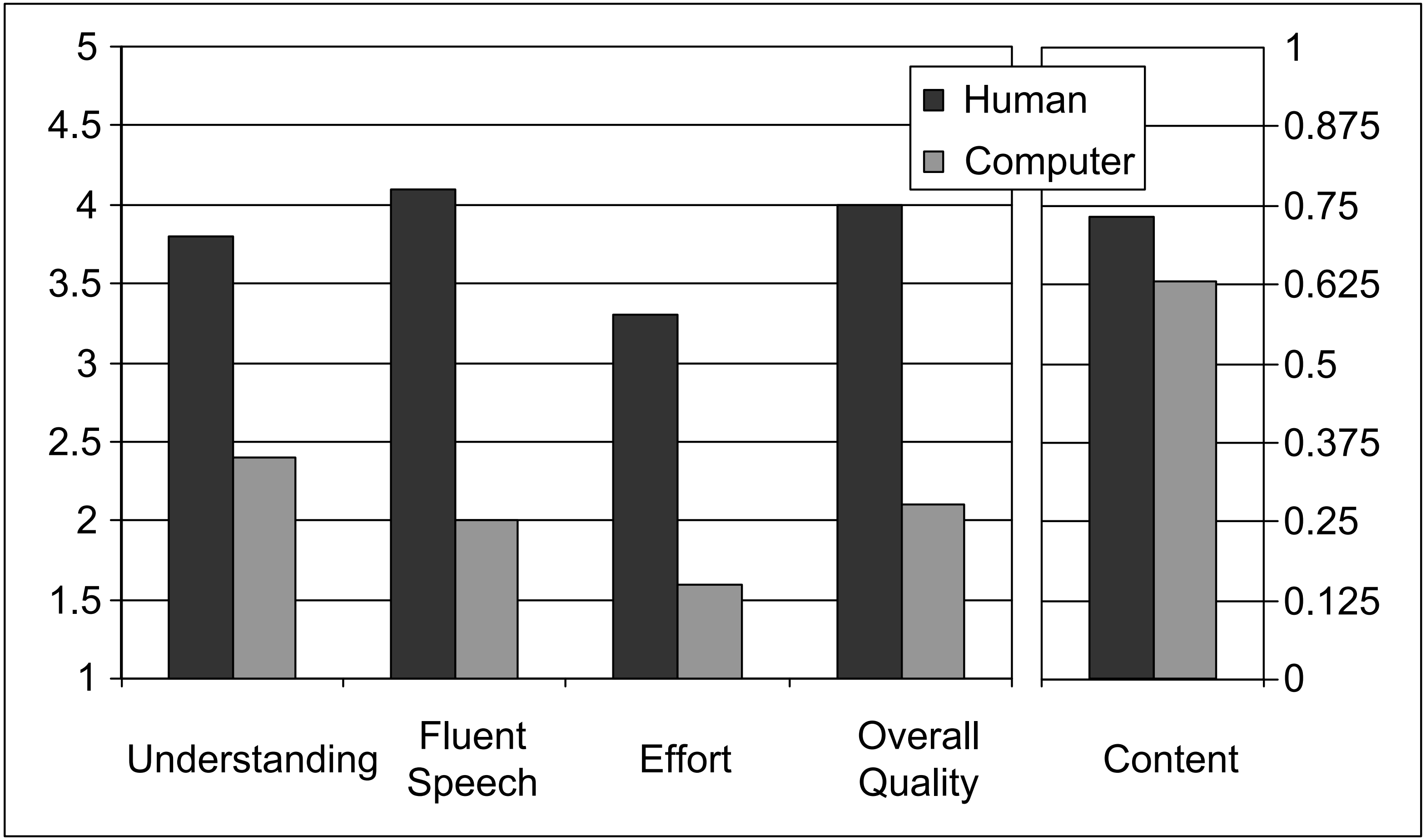
Approaches

- Sequence labeling
  - Input: words
  - Output: Labels

- Difficulties:
  - No word changes possible

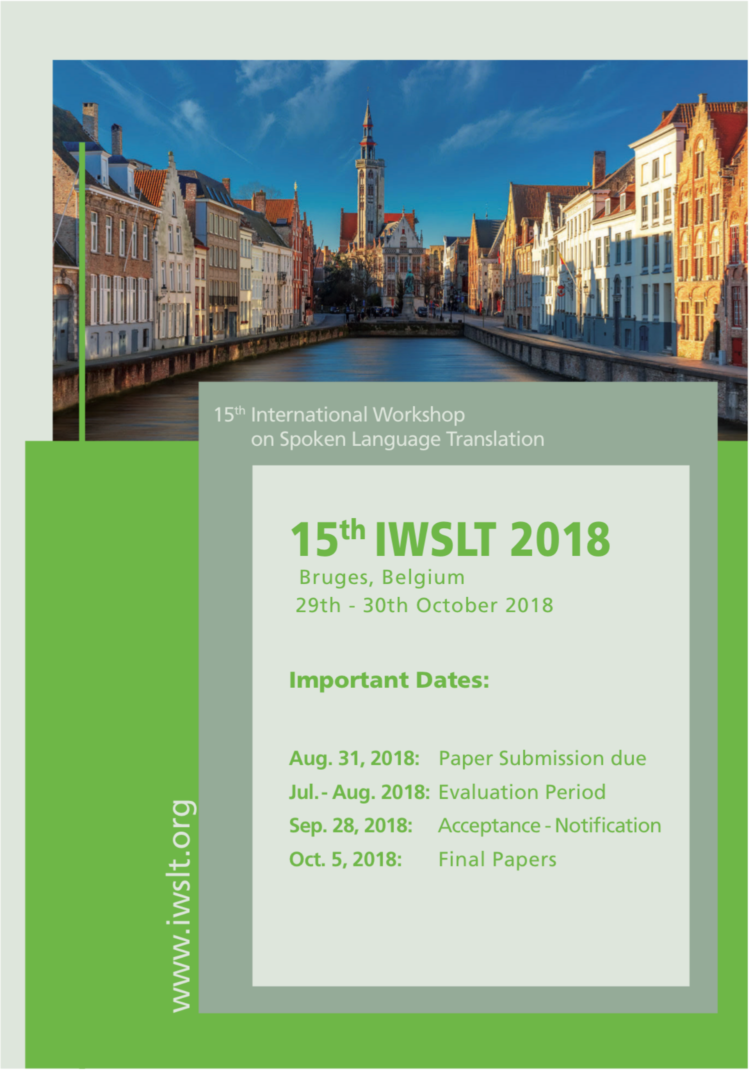


Human vs. Machine Performance



Summary

- Speech translation adds additional difficulties
  - Segmentation
  - Disfluencies
  - Latency
- Cascade models often still state of the art
- First successful applications
- Several scenarios still need research


15th International Workshop
on Spoken Language Translation
15th IWSLT 2018
Bruges, Belgium
29th - 30th October 2018
Important Dates:
Aug. 31, 2018: Paper Submission due
Jul. - Aug. 2018: Evaluation Period
Sep. 28, 2018: Acceptance - Notification
Oct. 5, 2018: Final Papers
www.iwslt.org