Teaching Machine Translation additional Constraints

Jan Niehues

23/06/2020

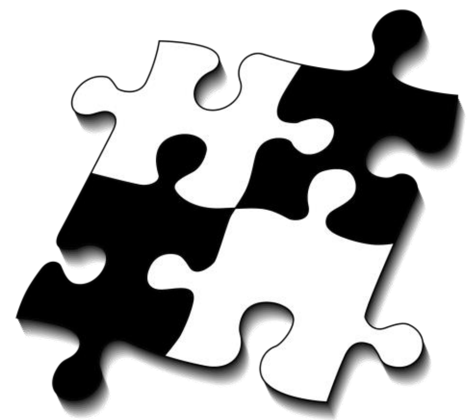
jan.niehues@maastrichtuniversity.nl

Maastricht University

Motivation

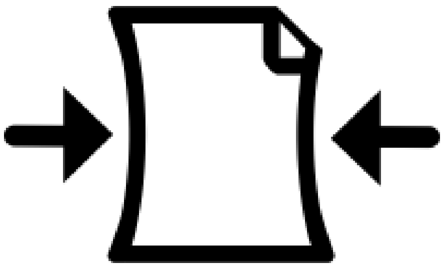
- NMT reach very good quality
  - Condition
    - Large amount of training data
    - Similar domain of training and test data

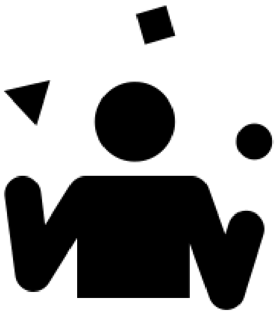
- Real-world applications
  - Often additional constrained necessary
    - Length constraints
    - Time constraints
  - No training data available



Length-constrained machine translation

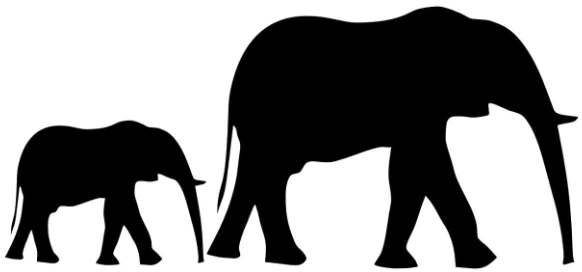
- Generate translation with a given length
  - Focus on shortening
- Translations of websites
  - Fit into layout
- Subtitles
  - Cognitive load
    - Adjust to reading speed





Time-constrained machine translation

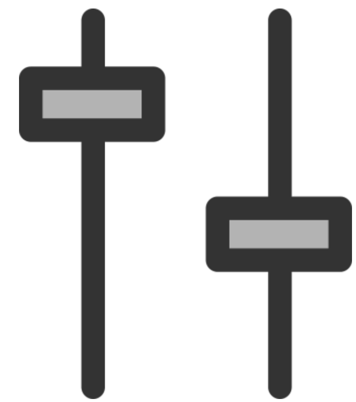
- Live transcription
  - Cannot wait for full sentence
- Strategies to output intermediate outputs
  - Update pervious outputs
  - Dynamically decide when to output
- Latency:
  - Time between spoken words and display of the translation



Overview

- Motivation
- Length Constraints
- Readability in subtitles
- Low-latency sequence-to-sequence models

Length-constrained translation

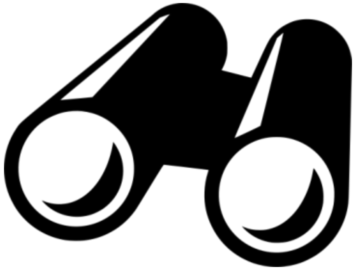


- Aim:
  - User is able to control length of translation
- Input:
  - Source language sentence
  - Desired target length
- Output:
  - Target sentence fulfilling length contained
    - Soft/Hard constraints
- Variants
  - Mono-lingual translation/Paraphrasing

Baseline

- Restrict search space
  - Only generate hypothesis fulfilling length constraint
  - Limit has research, increase probability of </s>
- Hard constraint
- No modification of training
- Problems:
  - Beginning of sentence cannot be changed

Idea

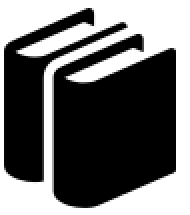


- Length aware during the whole generation
  - Plan your available spots
  - Shorten already at the beginning

- Challenges:
  - Target length also known during training
    - Training data with length
  - How to integrate length into model

Pseudo-supervised training

- Goal:
  - Training data with given target length
  - Available with different length ratios



- Challenge:
  - Hard to acquire

Source		Es klingt vielleicht übel.
Reference	8	It might sound like it's a bad thing.
Reference	4	It might sound bad.

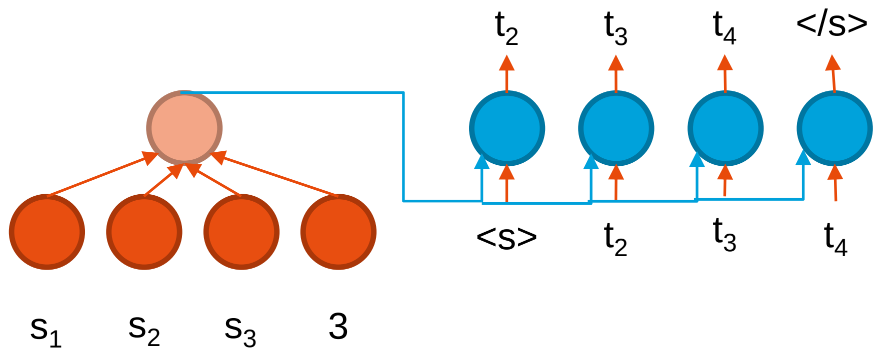
Pseudo-supervised training

- Idea:
  - Assume parallel training data was generated using length constraints
- Advantage:
  - Hugh amounts of training data for different domains
- Disadvantage:
  - Model might ignore length information

Source		Es klingt vielleicht übel.
Reference	8	It might sound like it's a bad thing.

Length representation

- Use length as additional input the encoder
  - Successfully done in multi-lingual MT, domain adaptation,…

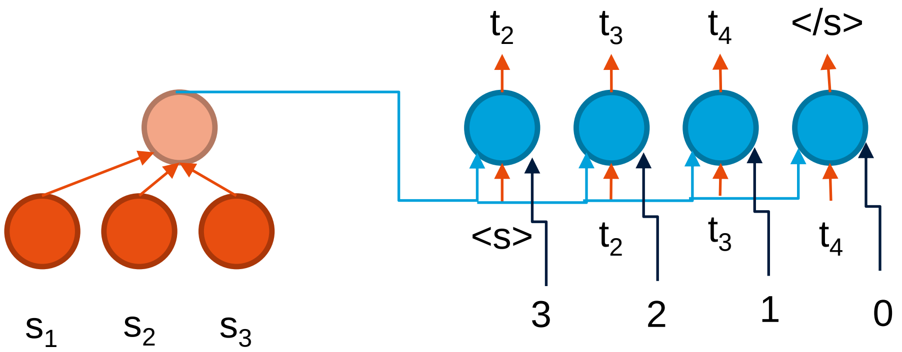


- Challenge:
  - Long lengths might be rare (e.g. 63)
  - Model might ignore length due to long distance from loss

Length representation

- Integrate into decoder
  - More direct influence on output probability

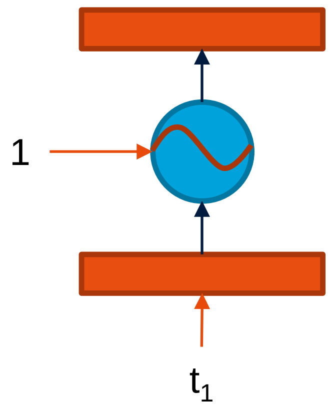
- Use remaining length at each step of the decoding process
  - Countdown to sentence end
  - Similar to positional encoding

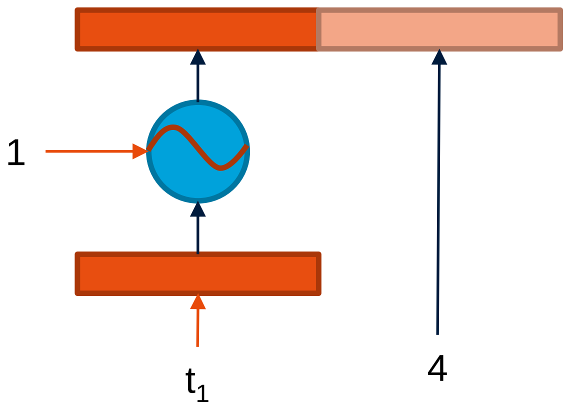


Length representation

- Include length information into the initial representation of each target work

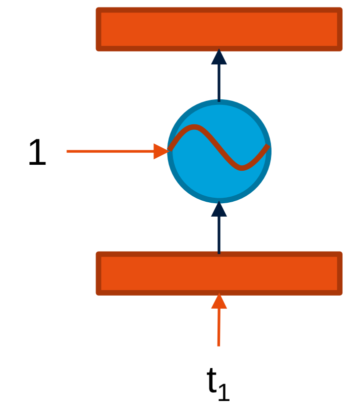
- Embedding
  - Concatenate embedding for the remaining length

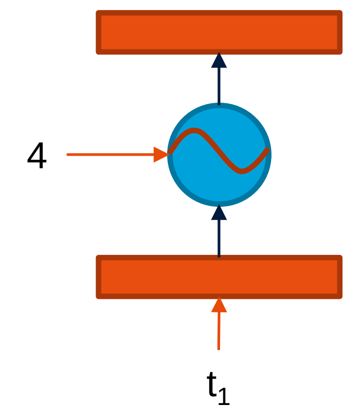




Length representation

- Include length information into the initial representation of each target work
- Embedding
  - Concatenate embedding for the remaining length
- Positional encoding
  - Encode remaining length instead of position





Evaluation

- No available evaluation data
- Use automatic metrics against original reference
- Problem:
  - Word-based metrics

Reference	It might sound like it's a bad thing.
Baseline	But it might sound like
Constraint	It sounds really bad .

- Embedding-based metric
  - RUSE

System

- IWSLT Multi-lingual set (2017)
  - German, English, Italian, Dutch and Romanian
  - Or German-English subset
  - Standard preprocessing with BPE
  - Target length: 80% and 50% of the source sentence
- Transformer
  - 8-layers
  - 512/2048

Task difficulty

- Force reference length

Model	BLEU	RUSE
Baseline	30.80	-0.085
Only Search	28.32	-0.124
Source Emb	28.56	-0.126
Decoder Emb	27.88	-0.140
Decoder Pos	28.80	-0.138

Length representation

- RUSE scores

Model	80%	50%
Baseline	-0.272	-0.605
Source Emb	-0.263	-0.587
Decoder Emb	-0.247	-0.555
Decoder Pos	-0.260	-0.577

Multi-lingual

- English-English as zero-shot translation of multi-lingual machine translation system
  - Target Length 80%

Model	Baseline	Decoder Emb.
DE-EN	-0.225	-0.214
EN-EN	-0.102	0.020

Cascade vs. End-to-End

- Target length 80%

Model	DE-EN	EN-EN
End2-End	-0.247	0.020
Cascade	-0.259	-0.118
Cascade Fix. Pivot		-0.166

Examples

Source	Und, obwohl es wirklich einfach scheint, ist es tatsächlich richtig schwer, weil es Leute drängt sehr schnell zusammenzuarbeiten.
Reference	And, though it seems really simple, it's actually pretty hard because it forces people to collaborate very quickly.
Base 0.8	and even though it really seems simple , it is actually really hard , because it really pushes
Dec. Emb 0.8	and although it really seems simple , it is really hard because it drives people to work together .
Base 0.5	and even though it really seems simple , it is really hard
Dec. Emb 0.5	it is really hard because it drives people to work together .

Simplification

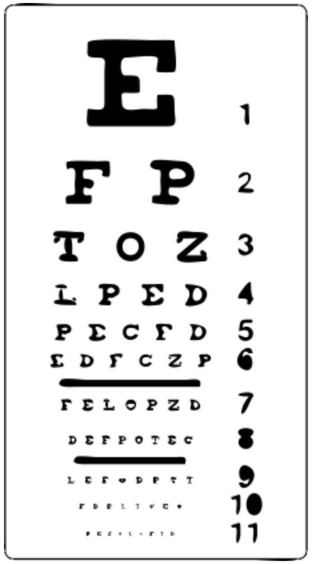
- Can we use the same framework for other tasks?
- Simplification:
  - Assumptions:
    - Words split by BPE are complex
    - Minimize split words
  - Approach:
    - Only count sub words
    - Generate translation with target length 0

Simplification - Result

Metric	Base	Simplified
BPE tokens	1899	991
DCI	7.66	7.45
BLEU	32.84	31.29

Readability

- Until now:
  - Compared to default translation
- Comparison to human subtitles
  - Generate for German TV News
- Monolingual
  - Aligned with audio



Example

"Befreit vom fraktionszwang soll das Parlament wohl nach der
"Free from party-constraints should the parliment maybe after the
Sommerpause die ethisch schwierige Frage debattieren."
summer break the ethically difficult question debate."

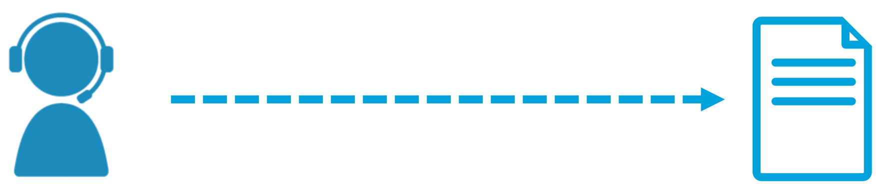
Example

"Befreit vom Ohne fraktionszwang soll das Parlament wohl nach der
"Free from Without party-constraints should the parliment maybe after the
Sommerpause die ethisch schwierige Frage debattieren."
summer break the ethically difficult question debate."

Experiments

- Casaded
  - **First** ASR
  - **Then** compression

- End-to-End
  - Transcription & compresion **in one model**



Datasets

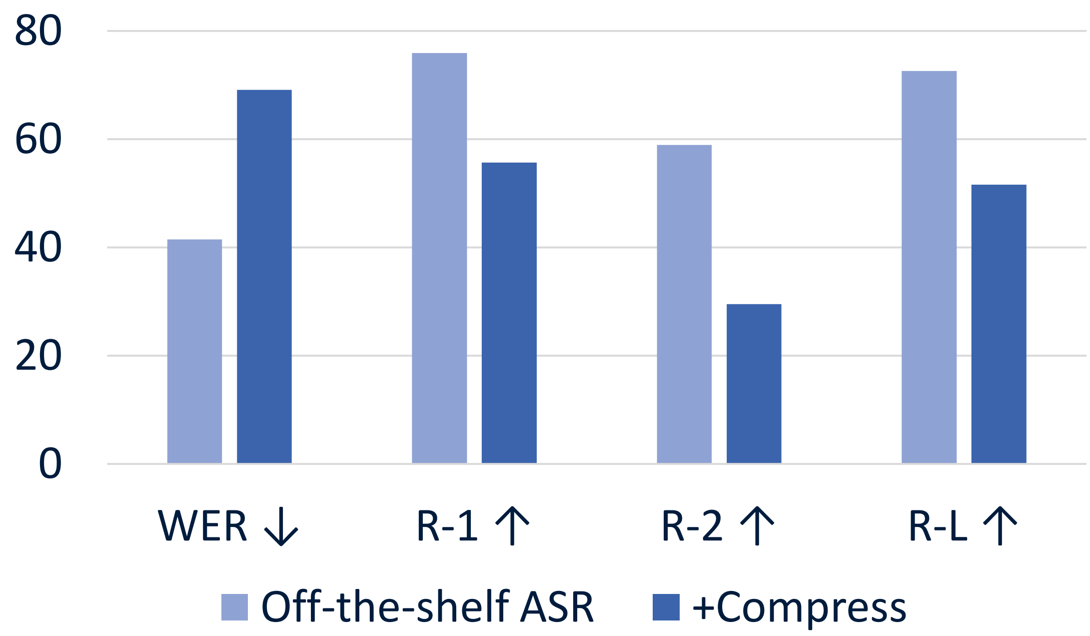
- Unsupervised compression model
  - {de, en, it, nl, ro} TED talks from IWSLT 2017

- End-to-end model

Partitions	Total length (h:m)	Total utterances
LibriVoxDeEn (**train**)	469:21	206,490
Tagesshau (**adapt**)	37:28	11,559
Tagesshau (**test**)	46	213

Cettolo et al. (2017). Overview of the IWSLT 2017 evaluation campaign. Proc. IWSLT.
Beilharz et al. (2020). LibriVoxDeEn: A Corpus for German-to-English Speech Translation and German Speech Recognition. Proc. LREC

Results



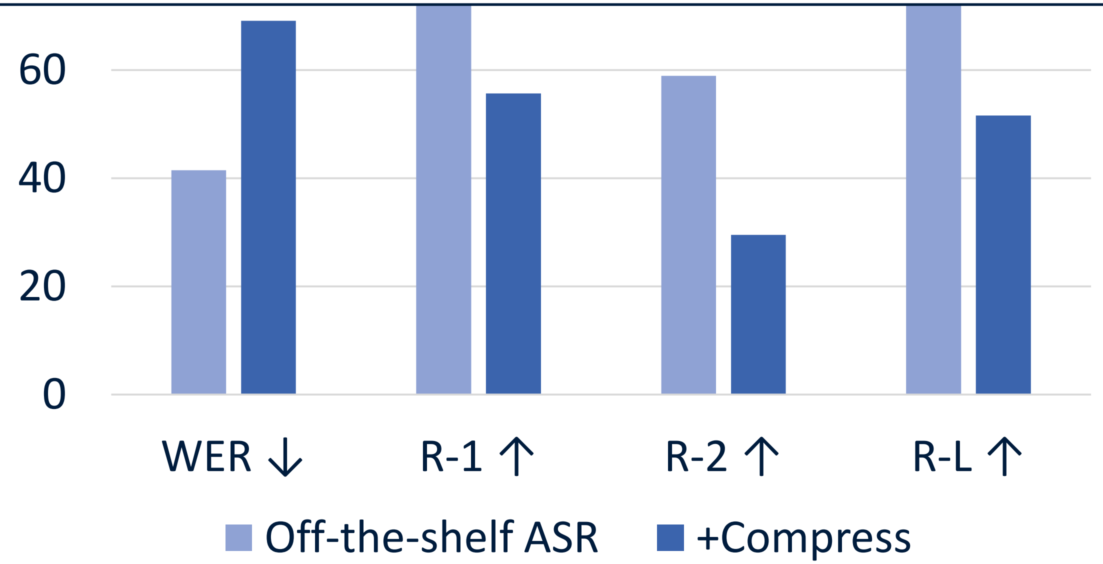
Results

Spoken: Es ist kurz nach Mitternacht, als plötzlich ein Auto in eine Gruppe von Menschen steuert, die ausgelassen ins neue Jahr feiern.

It is shortly after midnight, when suddenly a car into a group of people drives, that happily into new year celebrate.

Ref: kurz nach Mitternacht steuert ein Auto in eine Gruppe von Menschen , die ins neue Jahr feiern.

Output: kurz nach Mitternacht fährt ein Auto plötzlich in eine Gruppe von Leuten , die das nächste Jahr feiern.



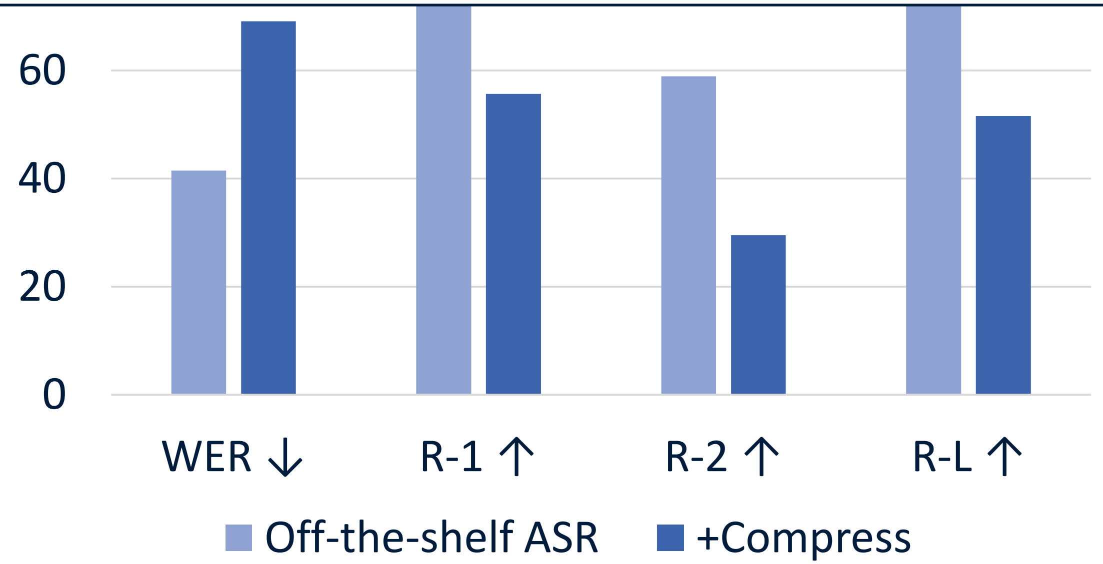
Results

Spoken: Es ist kurz nach Mitternacht, als plötzlich ein Auto in eine Gruppe von Menschen steuert, die ausgelassen ins neue Jahr feiern.

It is shortly after midnight, when suddenly a car into a group of people drives, that happily into new year celebrate.

Ref: kurz nach Mitternacht steuert ein Auto in eine Gruppe von Menschen , die ins neue Jahr feiern.

Output: kurz nach Mitternacht fährt ein Auto plötzlich in eine Gruppe von Leuten , die das nächste Jahr feiern.



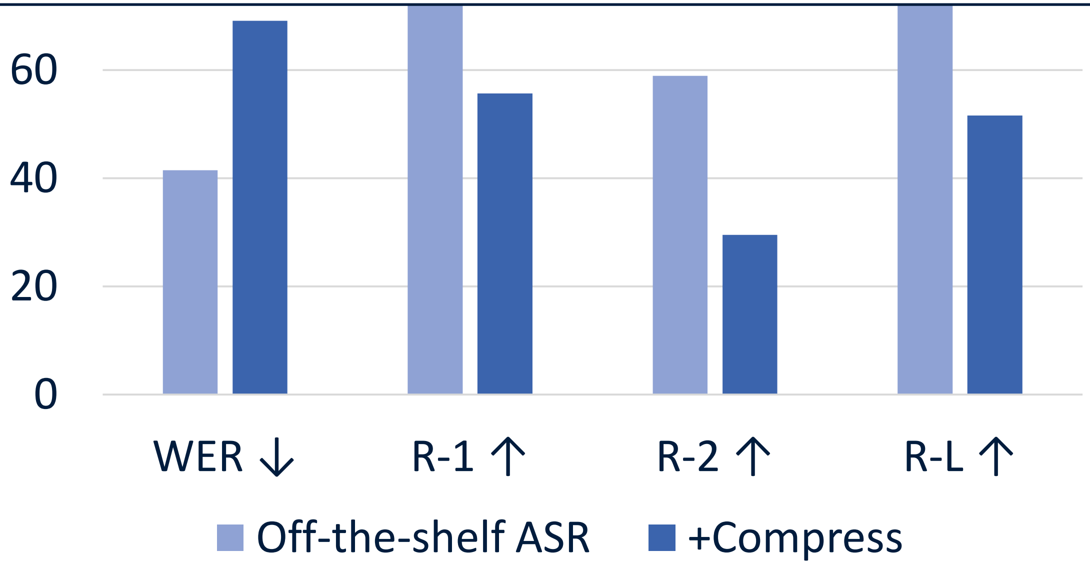
Results

Spoken: Es ist kurz nach Mitternacht, als plötzlich ein Auto in eine Gruppe von Menschen steuert, die ausgelassen ins neue Jahr feiern.

It is shortly after midnight, when suddenly a car into a group of people drives, that happily into new year celebrate.

Ref: kurz nach Mitternacht steuert ein Auto in eine Gruppe von Menschen , die ins neue Jahr feiern.

Output: kurz nach Mitternacht fährt ein Auto plötzlich in eine Gruppe von Leuten , die das nächste Jahr feiern.



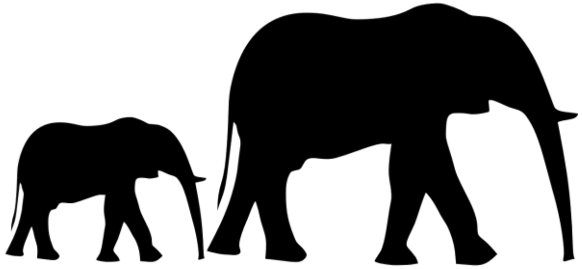
Explicit Length Constraints

- All models satisfy given length
- Encoding outperforms learned embedding
  - Unseen lengths in training

Adapted models	WER ↓	R-1 ↑	R-2 ↑	R-L ↑
Baseline (stop dec.)	39.9	74.6	**57.0**	72.6
Length embedding	39.3	74.3	55.2	72.5
Length encoding	**38.6**	**75.1**	56.4	**73.2**

Low-latency Sequence-to-Sequence Models

- Produce translation shortly after words are spoken
  - Before sentence ends
- Very short context
- Two techniques:
  - Iterative updates
  - Local agreement



Iterative Updates

- 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 [Niehues et al, 2016]

Adaptation to NMT

- Challenge:
  - NMT always tries to generate complete sentence
  - Example:
    - I encourage all of
    - Yo animo a todo el mundo .

Adaptation to NMT

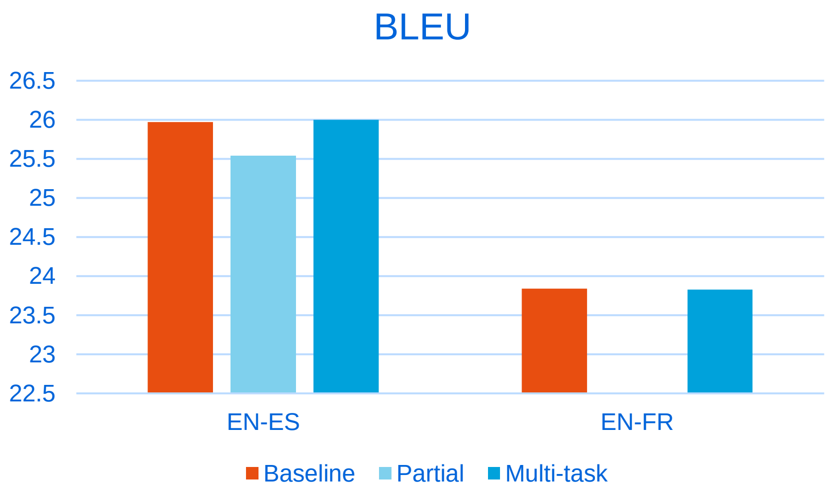
- Idea:
  - Train NMT on partial sentences
  - No parallel data available -> Generate artificial data

- Source data:
  - Every prefix of the sentence
- Target data:
  - Constraints:
    - As long as possible for low latency
    - Substring of previous prefix for few rewrites
  - Length-based
    - Same ratio of source and target sentence
  - Alignment-based:
    - Giza++ alignment
    - Longest prefix that no target word aligned outside source prefix

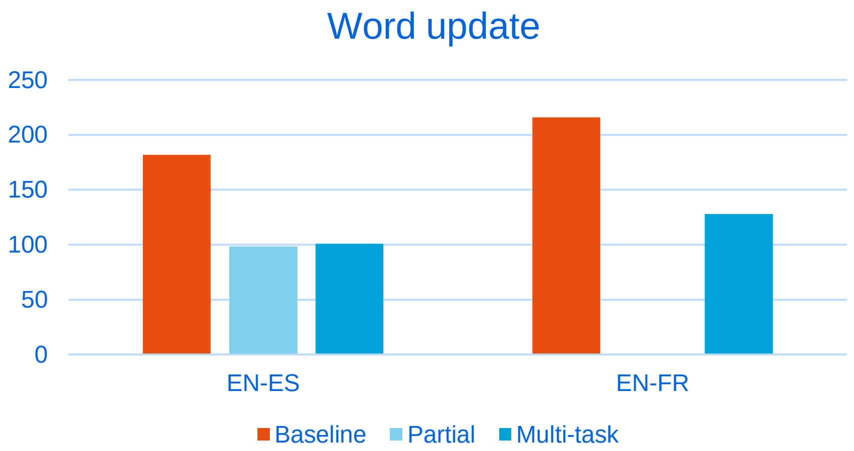
Adaptation to NMT

- Training
    - Continue training
        - Performance drop on full sentences
    - Multi-task training
        - Mix partial and full sentences
        - Ratio 1:1

Results



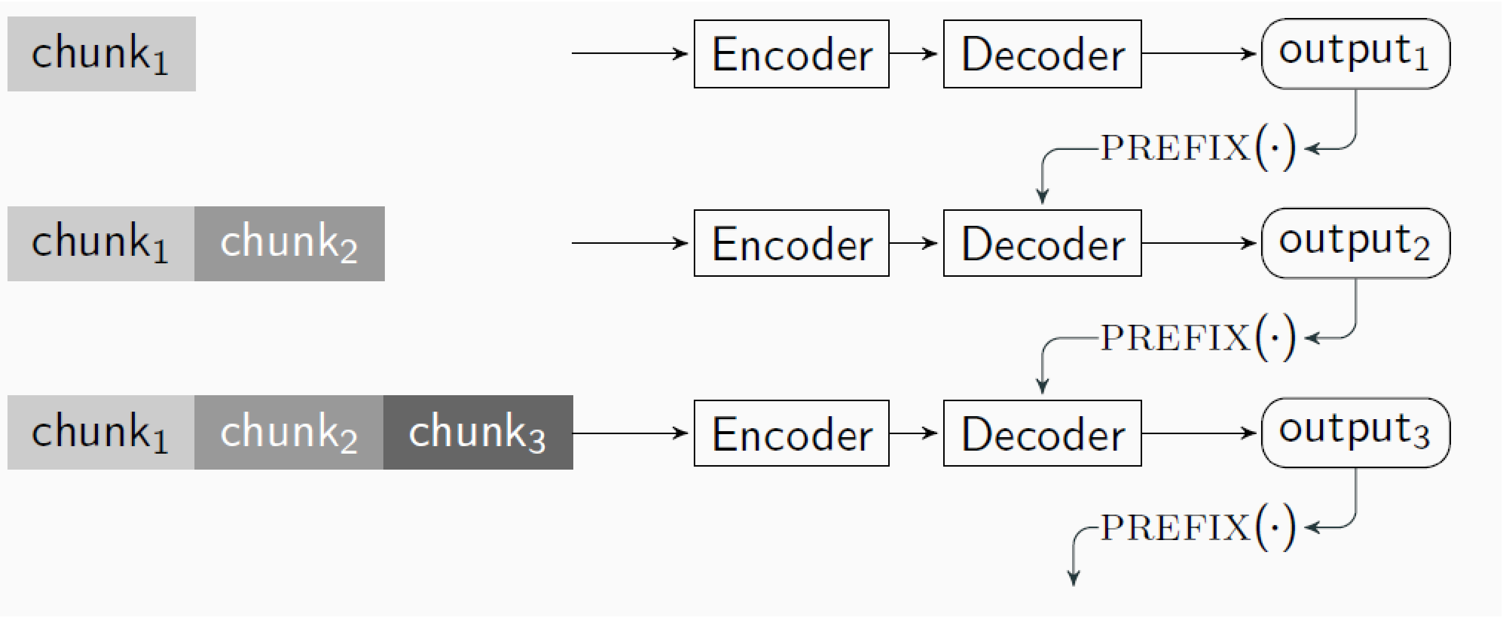
BLEU
26.5
26
25.5
25
24.5
24
23.5
23
22.5
EN-ES
EN-FR
Baseline
Partial
Multi-task



Word update
250
200
150
100
50
0
EN-ES
EN-FR
Baseline
Partial
Multi-task

Constrained Output

- Simulation framework
  - Evaluation different strategies



Stream decoding strategies

- Wait-k
  - Wait for k seconds
  - Then output with fixed rate

Chunks	Displayed	Output	Prefix
1	Ø	All model trains	Ø
1,2	Ø	All model art	All
1,2,3	All model	All models are ~~wrong~~	All model are
1,2,3,4	All model are		
…			

Stream decoding strategies

- Hold-n
  - Do not output last n tokens

Chunks	Displayed	Output	Prefix
1	Ø	All model ~~trains~~	All model
1,2	All model	All model ~~art~~	All model
1,2,3	All model	All model are ~~wrong~~	All model are
1,2,3,4	All model are		
…			

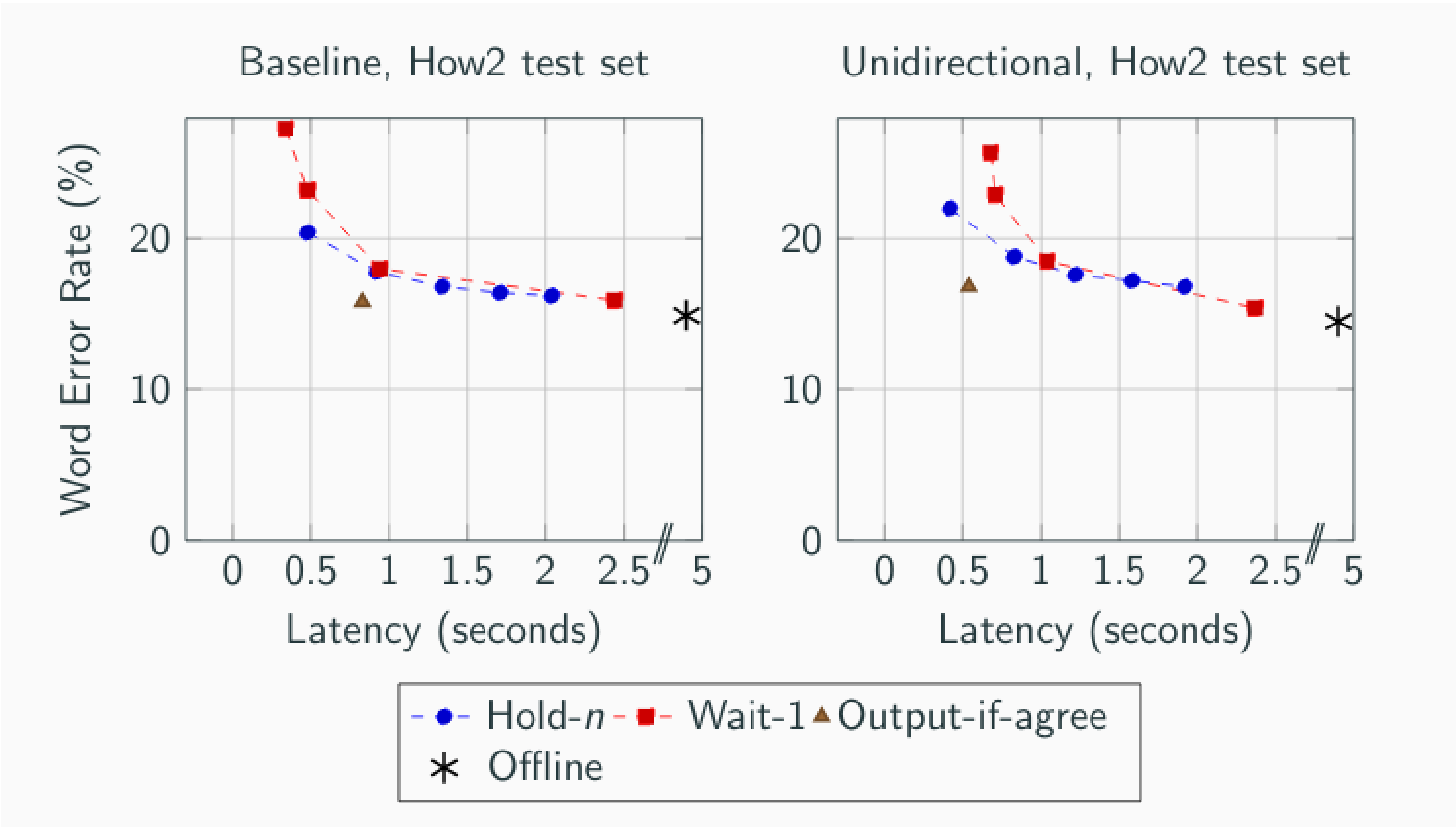
Stream decoding strategies

- Local agreement
  - Output if previous and current output agree on prefix

Chunks	Displayed	Output	Prefix
1	Ø	All model trains	Ø
1,2	Ø	All models art	All
1,2,3	All	All models are wrong	All models
1,2,3,4	All models		
…			

Latency vs. Accuracy

- Speech recognition results



Adaptation

- Adaptation to partial sentences:
  - Train on full and partial sentences

	Unidirectional	Bi-directional
Offline	14.4	14.9
Local agreement	16.8	15.8
+Adapt	15.5	15.8

Speech Translation

	BLEU	Latency diff.
Offline	44.5	4.36
Hold-2	37.3	0.48
Hold-4	42.2	0.95
Local Agreement	42.1	0.71

Conclusion

- Integration of additional constraints in NMT
  - Length-constraints
  - Time-constraints

- Architectural changes
- Pseudo-supervised training

- Length-constraints
  - Compared to human subtitles

- Time-constraints
  - Local agreement

Reference

- Niehues, J. (2020). *Machine Translation with Unsupervised Length-Constraints*. https://arxiv.org/pdf/2004.03176.pdf
- Liu, D., Niehues, J., & Spanakis, G. (2020). Adapting End-to-End Speech Recognition for Readable Subtitles. *Proceedings of the 17th International Workshop on Spoken Language Translation (IWSLT 2020)*.
- Niehues, J., Nguyen, T.-S., Cho, E., Ha, T.-L., Kilgour, K., Müller, M., Sperber, M., Stüker, S., & Waibel, A. (2016). Dynamic Transcription for Low-latency Speech Translation. *Proceedings of the 17th Annual Conference of the International Speech Communication Association (Interspeech 2016)*, 2513–2517. http://isl.anthropomatik.kit.edu/pdf/Niehues2016.pdf
- Niehues, J., Pham, N.-Q., Ha, T.-L., Sperber, M., & Waibel, A. (2018). Low-Latency Neural Speech Translation. *Proceedings of the 19th Annual Conference of the International Speech Communication Association (Interspeech 2018)*. https://www.isca-speech.org/archive/Interspeech_2018/pdfs/1055.pdf
- Liu, D., Spanakis, G., & Niehues, J. (2020). *Low-Latency Sequence-to-Sequence Speech Recognition and Translation by Partial Hypothesis Selection*.

Thanks