Data with Linguistic Knowledge for NMT

Part of the EAMT 2024 Tutorial

Linguistically Motivated Neural Machine Translation

Haiyue Song

https://shyyhs.github.io

European Association for Machine Translation

NICT National Institute of Information and Communications Technology

Roadmap

- Linguistic knowledge in

① **Data** Pre-Processing  e.g. watch/ing ab/normal/ly

current

② **Model** Training

③ **Decoding**

④ **Evaluation**

Checklist

- ☑ Ambiguity
- ☑ Composition
- ☒ Punctuation
- ☑ Verb tense
- …

Inject Linguistic Knowledge into Training Data

Data *Tokenization*

Word **Segmentation** [1]

Japanese 私/も/あさって/日曜/最終/日

I / also / day after tomorrow / Sunday / last / day

Subword **Segmentation** [2][3]

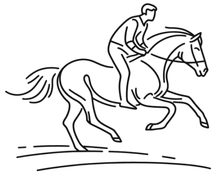
watch/ing sea/side

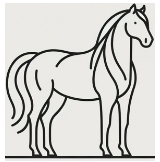
ab/normal/ly save/r/s

Character **Decomposition** [4]

Chinese 驰 → 马/也

run horse





Data from *Related Languages*

Script **Mapping** [5]

Zh: 今日天气晴，适合出门散步

Mapping:

Synthetic Ja: 今日天気晴，適合出門散步

Script **Romanization** [6]

Chinese 她到塔皓湖去了

……

Romanization ta dao ta hao hu qu le

She went to Lake Tahoe.

Word Segmentation: Motivation

- Add word boundary ⇨ **less ambiguity**

Japanese*

外国人参政権

right of foreigners to vote

✓ 外国 / 人 / 参政 / 権

foreign / people / suffrage / right

✗ 外国 / 人参 / 政権

foreign / carrot / regime

- Add word boundary ⇨ **better alignment**

Chinese**	目前出现与微信、支付宝结合的趋势
word seg.	目前 / 出现 / 与 / <u>微信</u> / 、 / <u>支付宝</u> / <u>结合</u> / 的 / <u>趋势</u>
English	There is a **trend** of **integration** with **WeChat** and **Alipay**

* Juman++: A Morphological Analysis Toolkit for Scriptio Continua

** PKUSEG: A Toolkit for Multi-Domain Chinese Word Segmentation

Word Segmentation: Example

- Segmentation results comparing
  - Juman++*

外国人参政権 right of foreigners to vote	→	外国 / 人 / 参政 / 権 foreign / people / suffrage / right
東京都知事 Tokyo governor	→	東京 / 都 / 知事 Tokyo / prefecture / governor

  - SentencePiece**

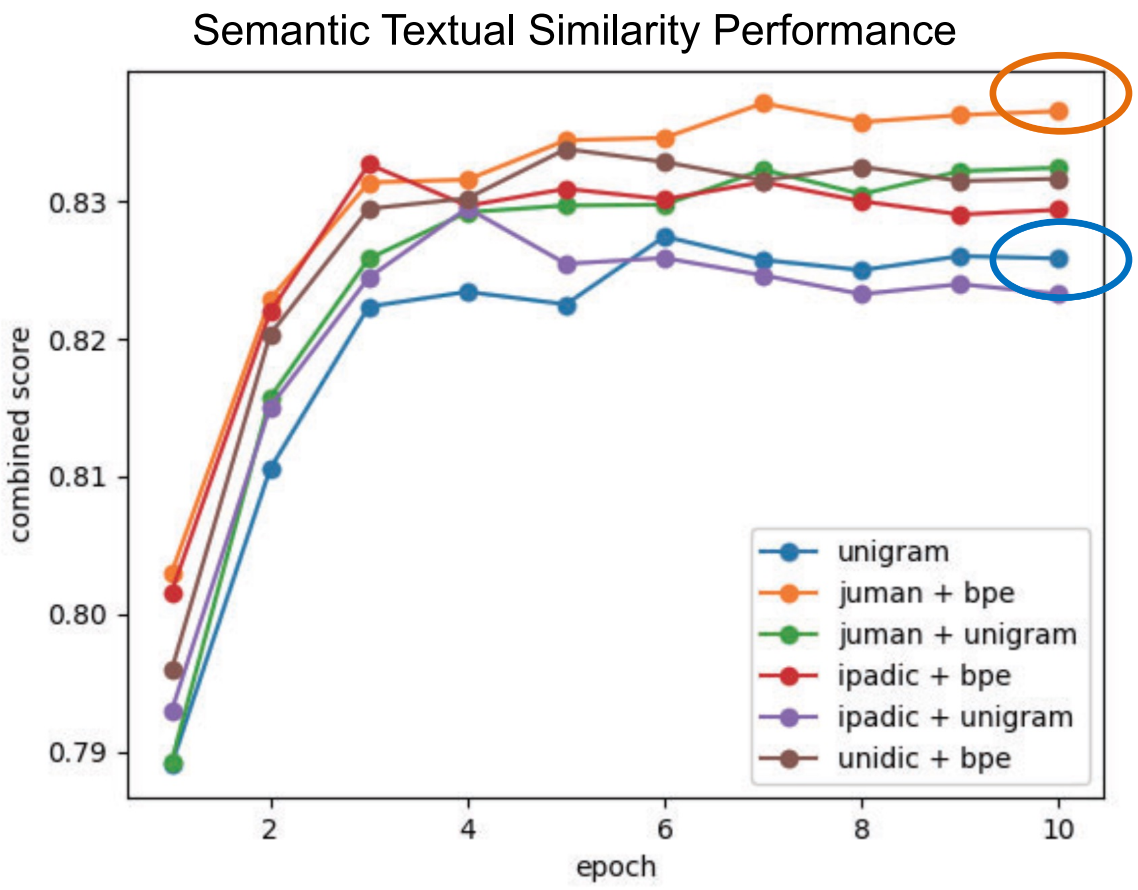
外国人参政権	→	外国 / 人 / 参 / 政権 foreign / people/ attend / regime
東京都知事	→	東 / 京都 / 知事 east / Kyoto/ governor

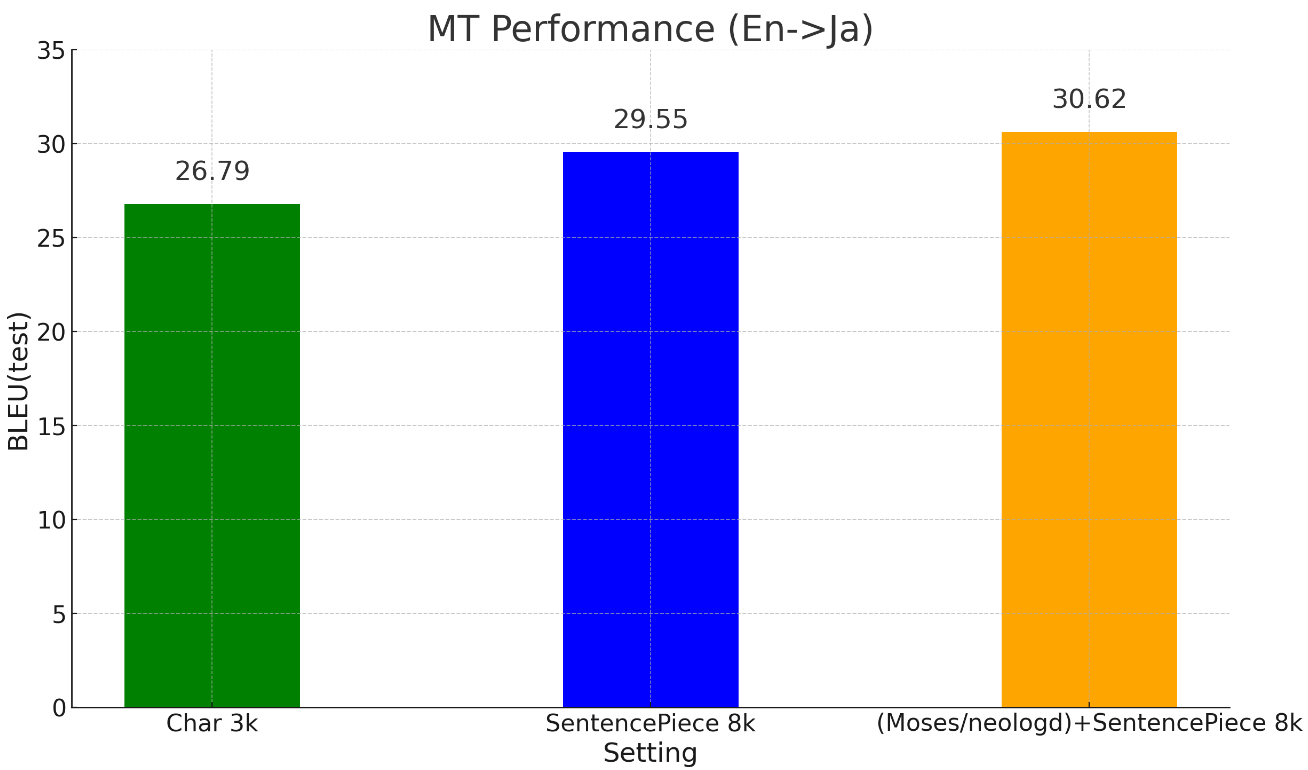
* Juman++: A Morphological Analysis Toolkit for Scriptio Continua

** SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing

Word Segmentation: Improvements

- Downsteam task performance
  - **Juman + bpe** > **SentencePiece-Unigram**





Subword Segmentation: Motivation

- NMT systems use **subwords** as the minimal unit.
- Compared to *word*, subwords handles **unseen words** by segmenting them into **seen subwords** in the subword vocabulary.

Sentence:	There are some trademarks.
Word segmentation:	There are some <UNK>.
Subword segmentation:	There are some trade_mark_s.

*<UNK> denotes *unknown words, which are not seen in the training corpora*

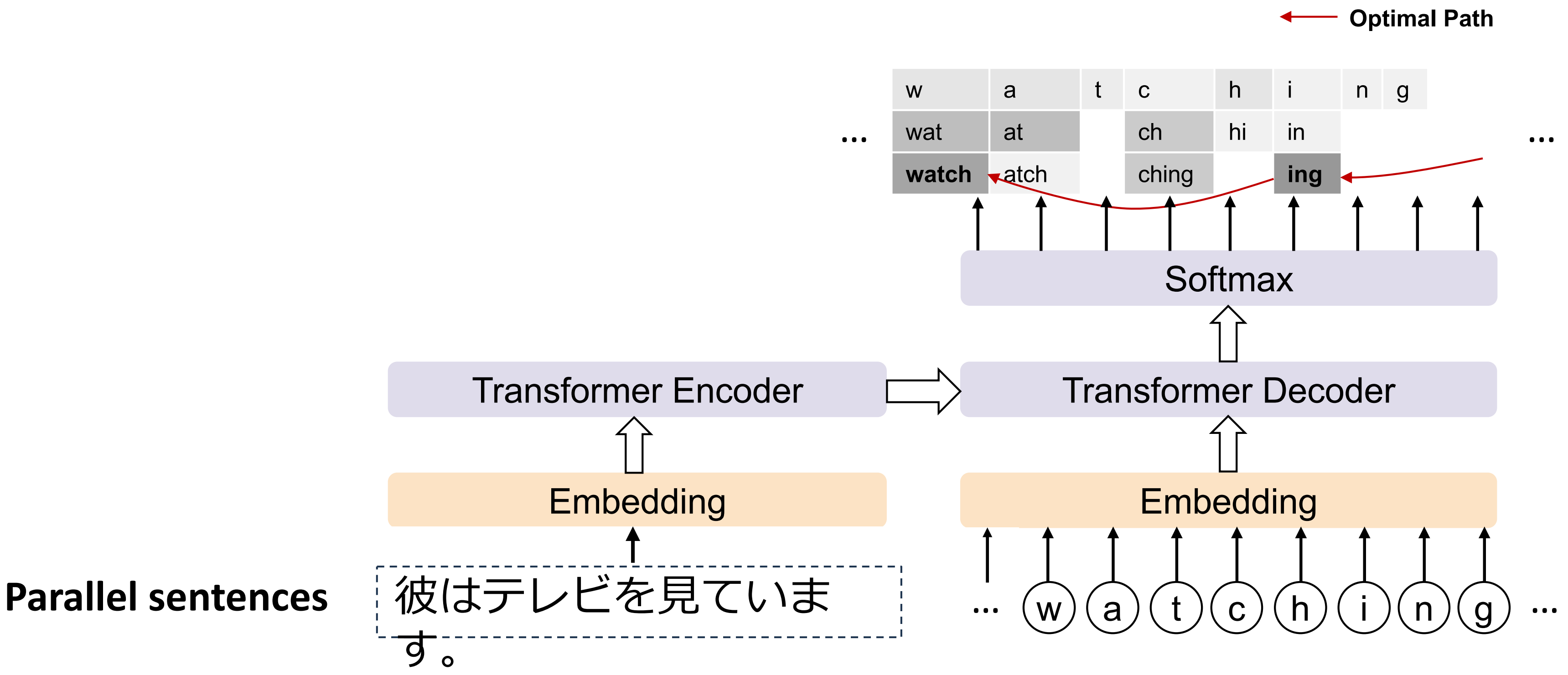
Linguistically Motivated Subword Segmentation

- Challenge: there are multiple segmentations for one word, which is the optimal one?

Words		**Frequency-based** (Sennrich+ 16)	**Linguistically Motivated** (Song+ 22, Song+ 23)
watching		wat + ching	watch + **ing**
languages		langu + ages	language + **s**
unknown	Segmentation →	un + k + n + own	**un** + know + **n**
seaside		se + as + ide	sea + **side**
abnormally		ab + n + orm+ ally	**ab** + normal + **ly**
savers		sa + vers	save + **r** + **s**

Subword Segmentation: Method

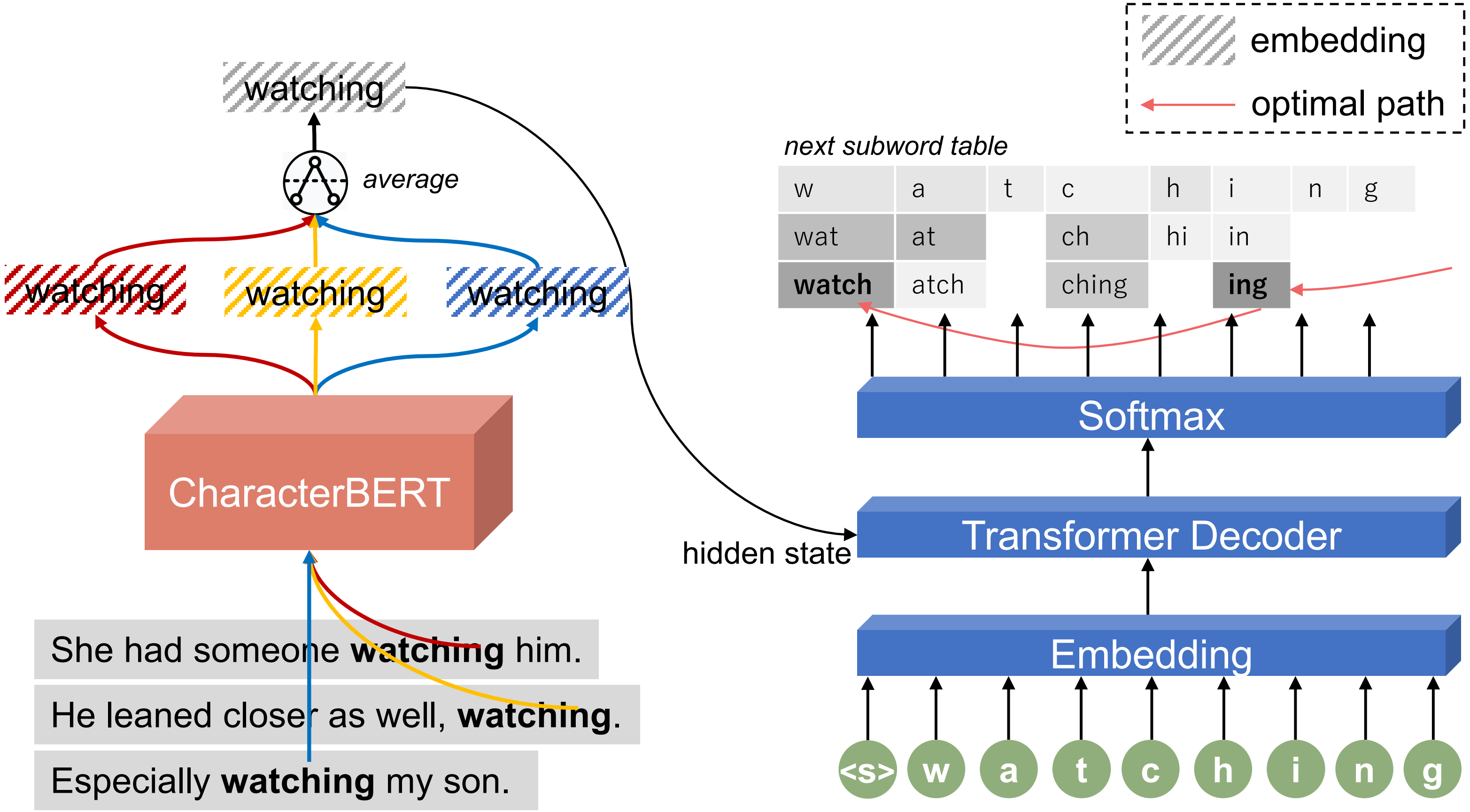
- Dynamic Programming Encoding (DPE) (He+ 20) is a neural segmenter trained on **parallel sentences**.
  - It maximizes the marginal likelihood of the **target sentence**.



[2] Dynamic Programming Encoding for Subword Segmentation in Neural Machine Translation

Subword Segmentation: Method

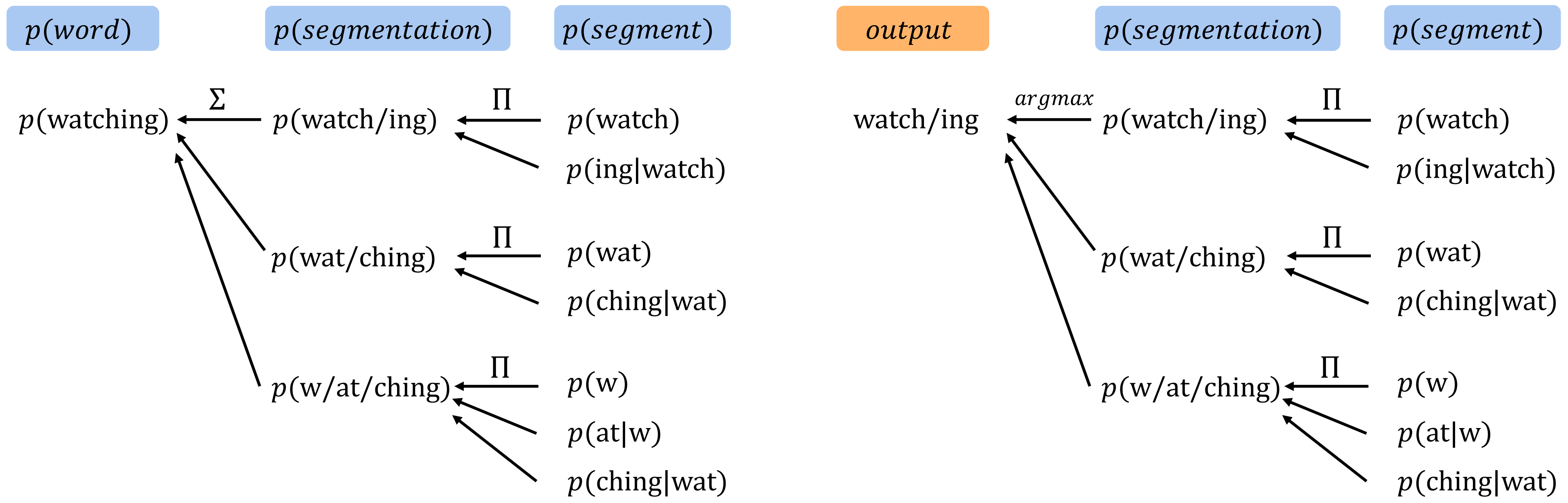
- BERTSeg
  - Uses semantic information from BERT embeddings.
  - It maximizes the marginal likelihood of the **target word**.



[3] BERTSeg: BERT Based Unsupervised Subword Segmentation for Neural Machine Translation

Subword Segmentation: Training and Decoding

- Training
  - Maximize the generation probability of word by maximizing all possible segmentations, *conditioned on its semantic embedding*.
- Decoding
  - Retrace the optimal segmentation with the maximum generation probability.
  - Stochastic version: sample segmentations based on their probabilities.



Subword Segmentation

- BERTSeg

BERTSeg	BPE
Frequent words	
official/s	officials
edit/ion	edition
use/d	used
farm/er/s	far/mers
contribute/d	contrib/uted
normal/ly	norm/ally
seven/th	sevent/h
challenge/d	challeng/ed
over/night	o/vern/ight
language/s	langu/ages

BERTSeg	BPE
Rare words	
inter/face/s	inter/f/aces
sea/side	se/as/ide
ab/normal/ly	ab/n/orm/ally
b/y/stand/er	by/st/ander
dis/comfort	disc/om/fort
un/warrant/ed	un/w/arr/anted
in/definitely	ind/ef/in/itely

BERTSeg	BPE
Unseen words	
stable/d	st/ab/led
save/r/s	sa/vers
M/illion/s	Mill/ions
Free/way	Fre/ew/ay
M/i/s/behavior	M/is/be/hav/ior
m/o/u/r/n/ed	m/our/ned
M/a/d/a/m/e	Mad/ame

Segmentation for Other Languages

- Use multilingual BERT encoder

Japanese

Word		Subwords
行った		行 + った
可能であった		可能 + であった
紹介	segmenter →	紹介
利用		利用

Chinese

Word		Subwords
优先度		优先 + 度
攻击者		攻击 + 者
独立性	segmenter →	独立 + 性
问题点		问题 + 点

Malay

Word		Subwords
bertanggungjawab responsible		ber + tanggungjawab responsibility
kontraktor contractor		kontrak + tor contract
membawanya bring it	segmenter →	membawa + nya bring it
berpakaian dress up		ber + pakaian clothes

Myanmar

Word		Subwords
ကိုယ်ရေးအချက်အလက် personal information		ကိုယ်ရေး + အချက်အလက် private/personal information
အစည်းအဝေးခန်း meeting room		အစည်းအဝေး + ခန်း meeting room
ပြောထား said	segmenter →	ပြော + ထား say
ကမ်းလှမ်းထား offered		ကမ်းလှမ်း + ထား offer

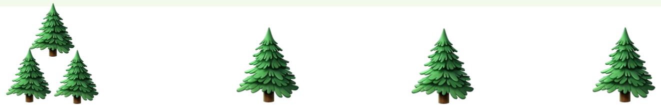
Character Decomposition: Motivation

- Characters in languages such as Chinese, Japanese, Korean may contain sub-characters.

Character Decomposition

森 → 木 / 木 / 木

forest tree tree tree



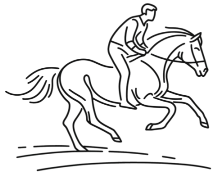
林 → 木 / 木

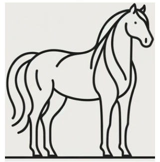
woods tree tree

Character Decomposition

驰 → 马 / 也

run horse





website for character decomposition: https://www.cns11643.gov.tw/searchQ.jsp?ID=0

Character Decomposition: Method

- Replacing characters with ideograph sequences in the training data.

	Language	Word
	JP-character	風 景
Method in this paper	JP-ideograph	几一虫 日亠口小_1
	JP-stroke	丿⺄一丨㇕一丨㇒丶 丨㇕一一丶一丨㇆一亅㇒丶_1
	CN-character	风 景
Method in this paper	CN-ideograph	几㐅 日亠口小_1
	CN-stroke	丿⺄㇒丶 丨㇕一一丶一丨㇆一亅㇒丶_1
	EN	landscape

Character Decomposition: Results

- Best performance compared to word/character/stroke

English-Chinese NMT		BLEU
EN_word	CN_word	11.8
EN_word	CN_character	10.3
EN_word	CN_ideograph	**14.6***
EN_word	CN_stroke	14.1*
Chinese-English NMT		**BLEU**
CN_word	EN_word	14.7
CN_character	EN_word	14.5
CN_ideograph	EN_word	**15.6***
CN_stroke	EN_word	15.5*

Using Data in Related Languages

- Motivation
  - Transfer knowledge in high-resource language to low-resource language
  - Especially helpful if they are related (share the same grammar etc.)
- Challenge
  - Often **different script**

Using Data in Related Languages

- From non-Latin to Latin

English

Big Ben

embedding

Well Learned

Romanization

Biggu Ben ← ビッグベン

Japanese

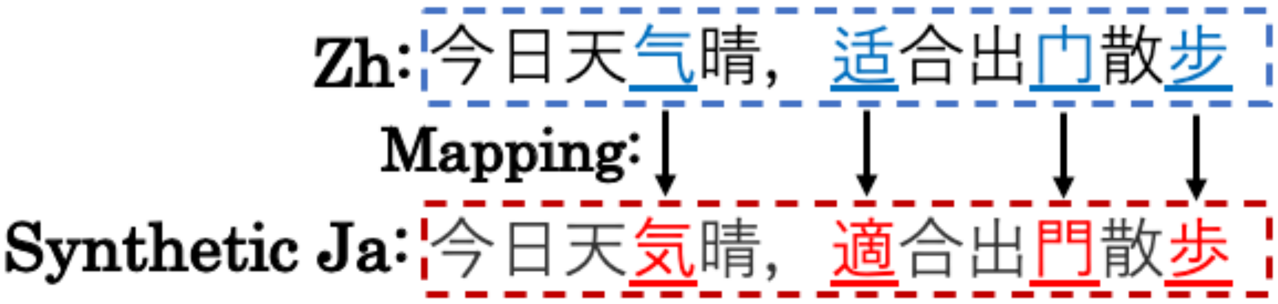
embedding

Poorly Learned

Using Data in Related Languages

- From one language to another **related** language

Script Mapping [5]



MT Performance using Romanization

- Improves the performance of **low-resource language** ↑
- But hurts the performance of **high-resource language** ↓

	base	transfer from multilingual parent		
		orig	uroman	uconv
am-en	14.4	**16.2**	**16.5**	**16.0**
en-am	12.7	13.7	6.5	**14.3**
mr-en	34.3	**45.0**	43.4	42.8
en-mr	25.7	**33.4**	**33.2**	**33.0**
ta-en	21.9	**29.3**	**29.0**	**29.2**
en-ta	13.5	**21.5**	21.0	**22.4**
avg imp	-	**+ 6.1**	+ 4.5	**+ 5.9**

	orig	uroman	uconv
ar-en	**37.4**	36.3	**37.4**
ru-en	33.3	33.5	**34.1**
zh-en	**39.5**	37.0	**39.2**

Summary

- Linguistic knowledge can be injected in the training data
  - Word segmentation for languages such a Japanese and Chinese
  - Linguistically motivated subword segmentation
  - Character decomposition
- Data from related languages helps through
  - Script mapping
  - Romanization

Open Questions

- Character-level/Byte level tokenization
  - Character-level contains all information in theory
  - But it underperforms subword based methods
  - Why? Because the current architecture is designed for subwords?
- Knowledge transfer only in similar script
  - e.g., if some knowledge appears in English, if the model outputs Japanese it can never access that knowledge
  - Romanization hurts the performance of the original language
  - how to transfer between different scripts efficiently?

References

[1] Juman++: A Morphological Analysis Toolkit for Scriptio Continua

[2] Dynamic Programming Encoding for Subword Segmentation in Neural Machine Translation

[3] BERTSeg: BERT Based Unsupervised Subword Segmentation for Neural Machine Translation

[4] Neural Machine Translation of Logographic Languages Using Sub-character Level Information

[5] Pre-training via Leveraging Assisting Languages for Neural Machine Translation

[6] On Romanization for Model Transfer Between Scripts in Neural Machine Translation

Linguistically Aware Decoding

Part of the EAMT 2024 Tutorial

Linguistically Motivated Neural Machine Translation

Haiyue Song

https://shyyhs.github.io

European Association for Machine Translation

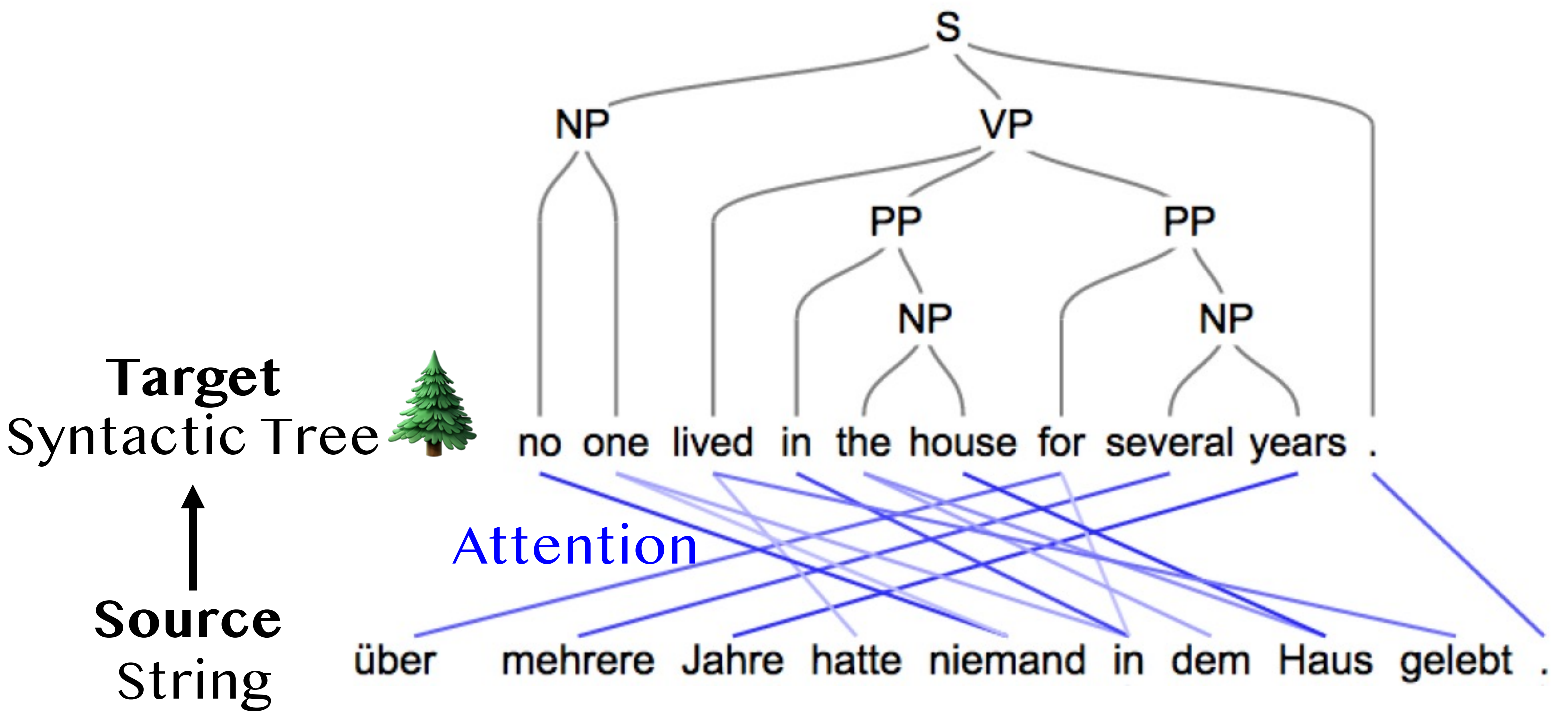
NICT National Institute of Information and Communications Technology

Decoding

- String-to-Tree Decoding
- Structural Template Prediction

String to Syntactic Tree

- Incorporate syntactic tree information [1]



[1] Towards String-to-Tree Neural Machine Translation

String to Syntactic Tree

- Syntactic tree is linearized for the NMT model to process
  - Syntactic info is obtained from a specific **English** parser

English **Jane had a cat .**

Linearized

Jane hatte eine Katze . $\rightarrow$ $(_{ROOT}$ $(_{S}$ $(_{NP}$ **Jane** $)_{NP}$ $(_{VP}$ **had** $(_{NP}$ **a cat** $)_{NP}$ $)_{VP}$ **.** $)_{S}$ $)_{ROOT}$

[1] Towards String-to-Tree Neural Machine Translation

Results: Helpful in Low-resource

- No large improvement in high-resource scenarios

4.5M parallel sentences
German➡️English

system	newstest2015	newstest2016
bpe2bpe	27.33	31.19
bpe2tree	27.36	32.13

166k parallel sentences low-resource scenario
German➡️English

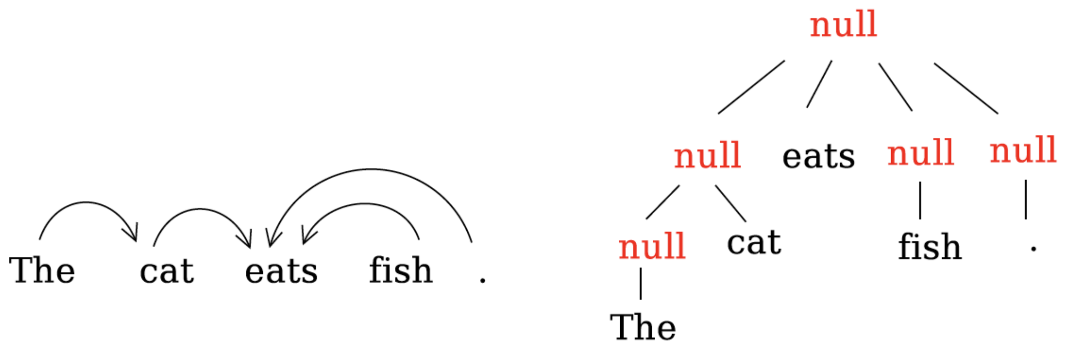
system	newstest2015	newstest2016
bpe2bpe	13.81	14.16
bpe2tree	14.55	16.13

[1] Towards String-to-Tree Neural Machine Translation

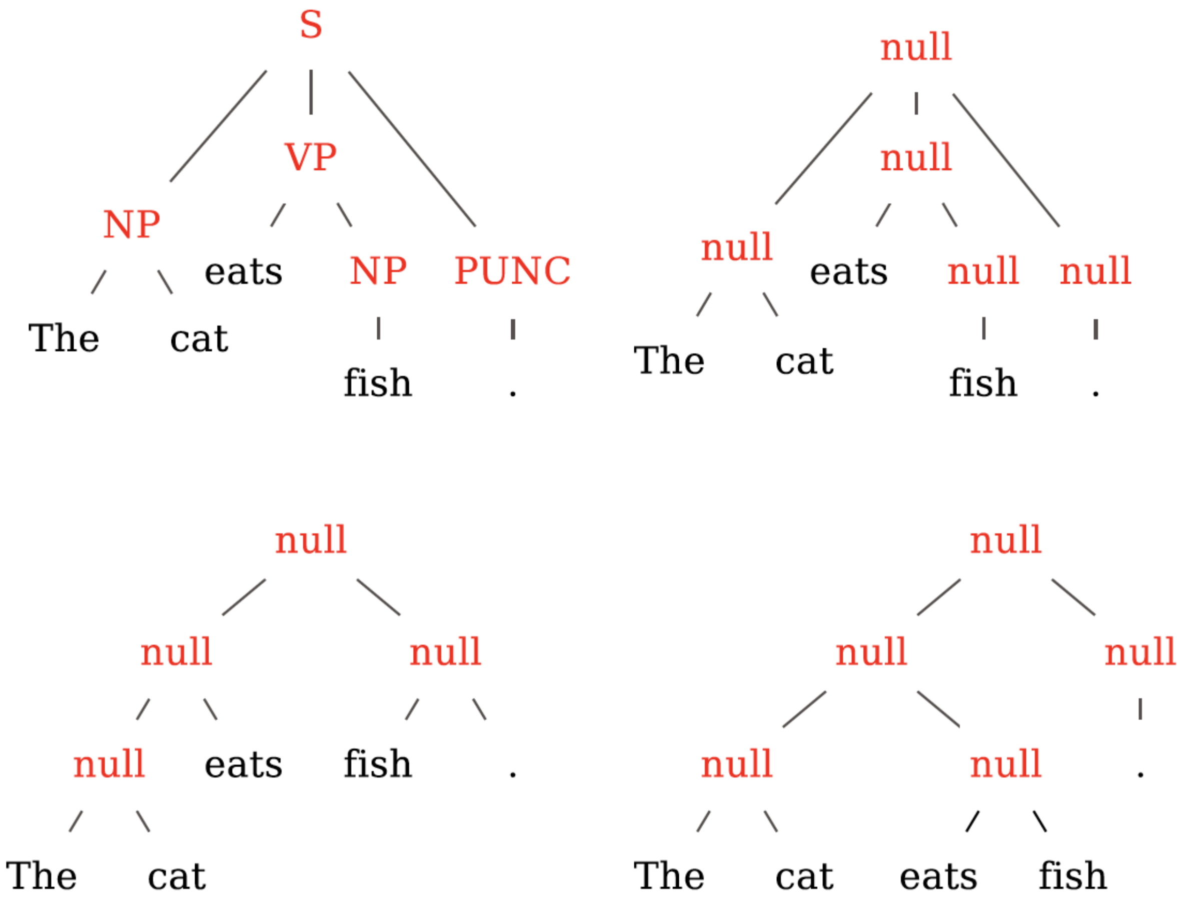
String to Any Tree

- Incorporate any tree structure information [2]

Example of one parser



Different parsers, or even not from parser



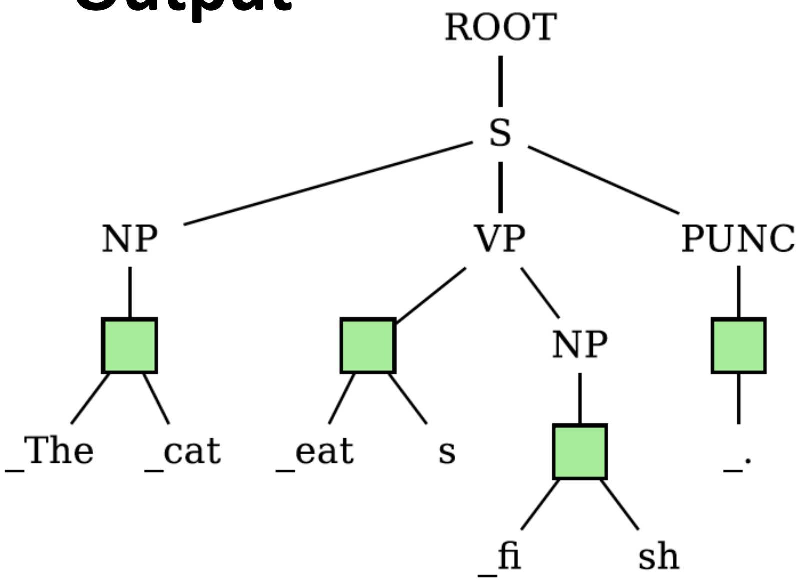
Balanced Binary Tree

[2] A Tree-based Decoder for Neural Machine Translation

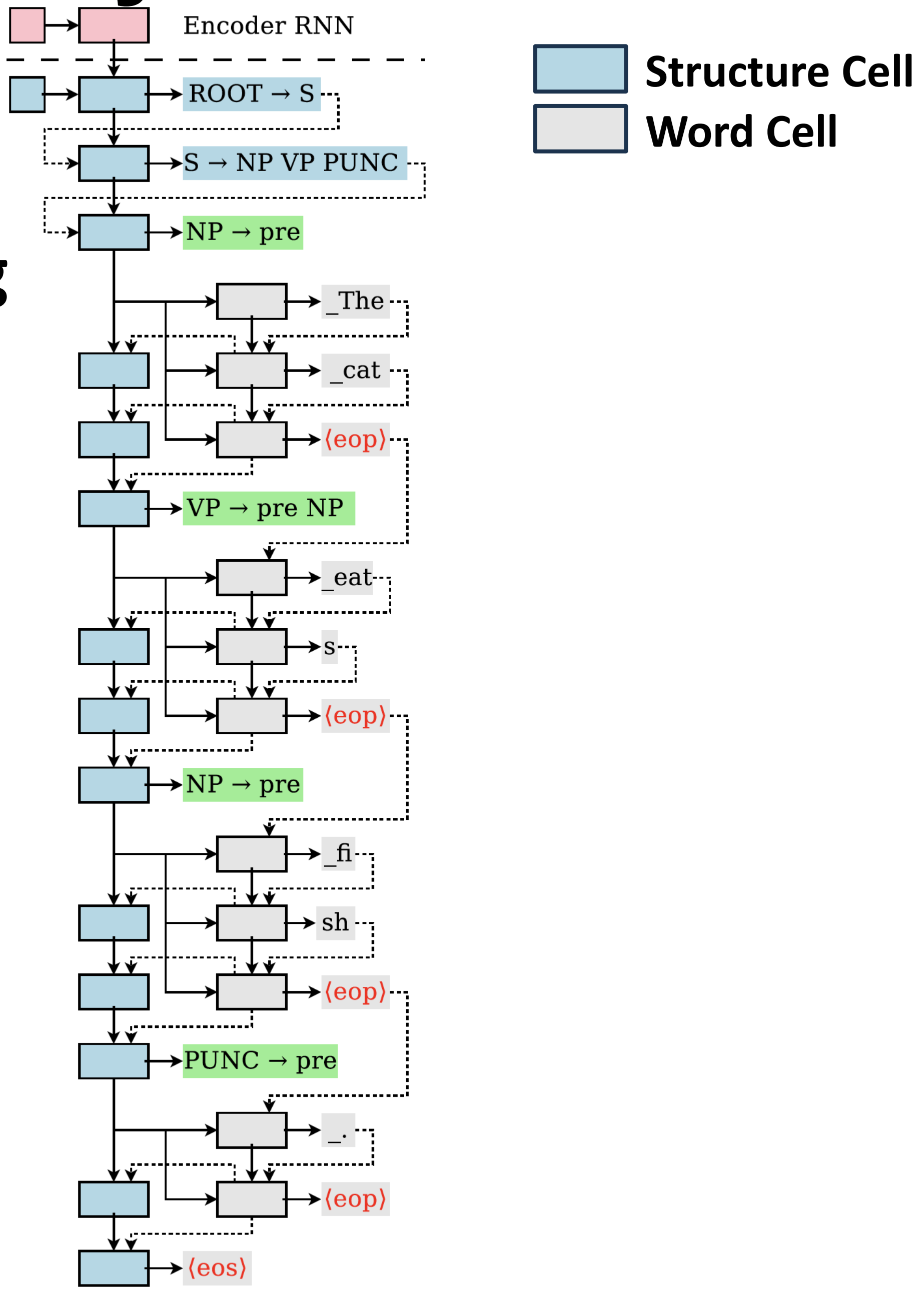
String to Any Tree

- Tree-structure-aware decoder [2]

Output



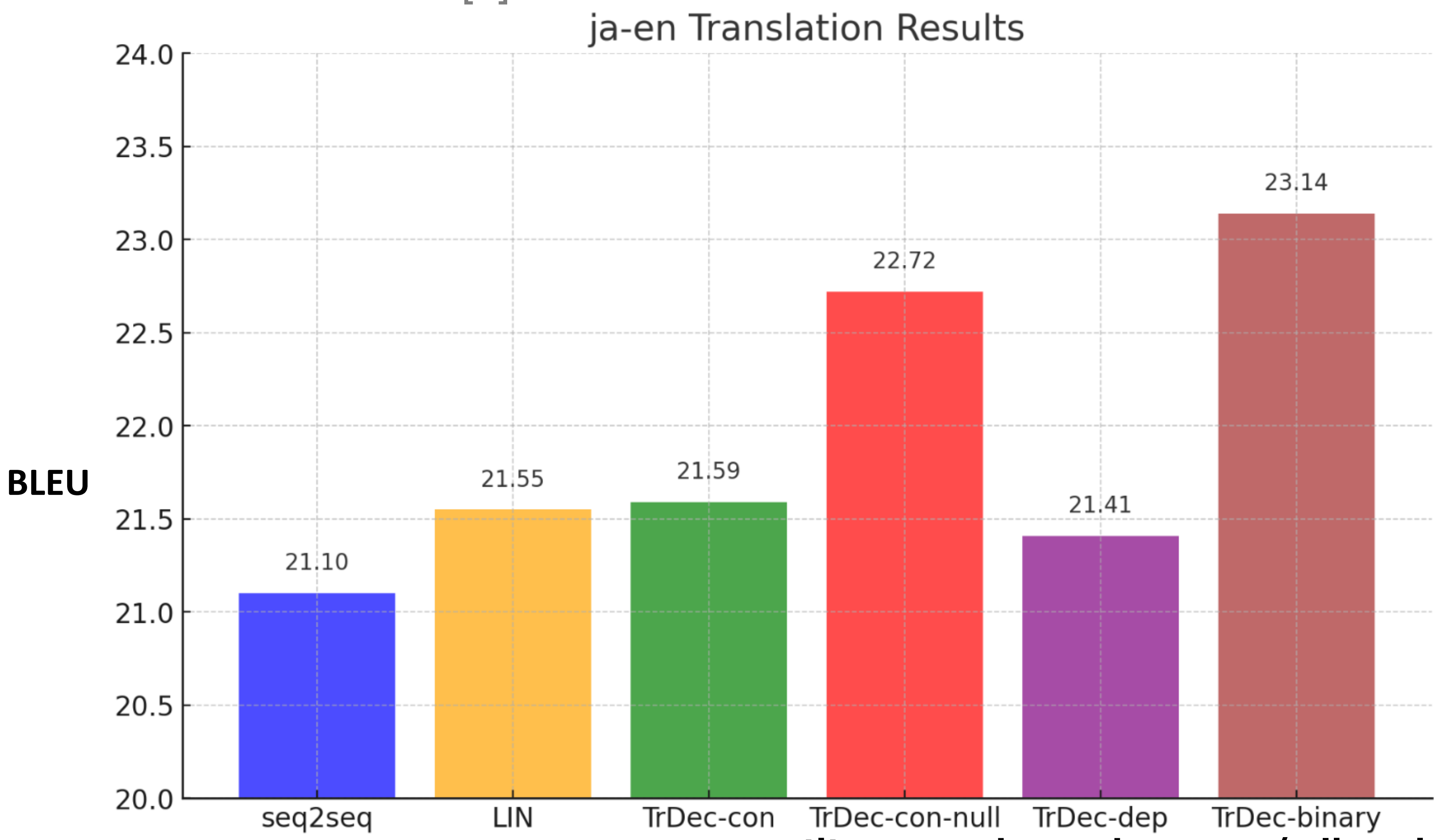
Decoding



[2] A Tree-based Decoder for Neural Machine Translation

String to Any Tree

- Balanced Tree works [2]

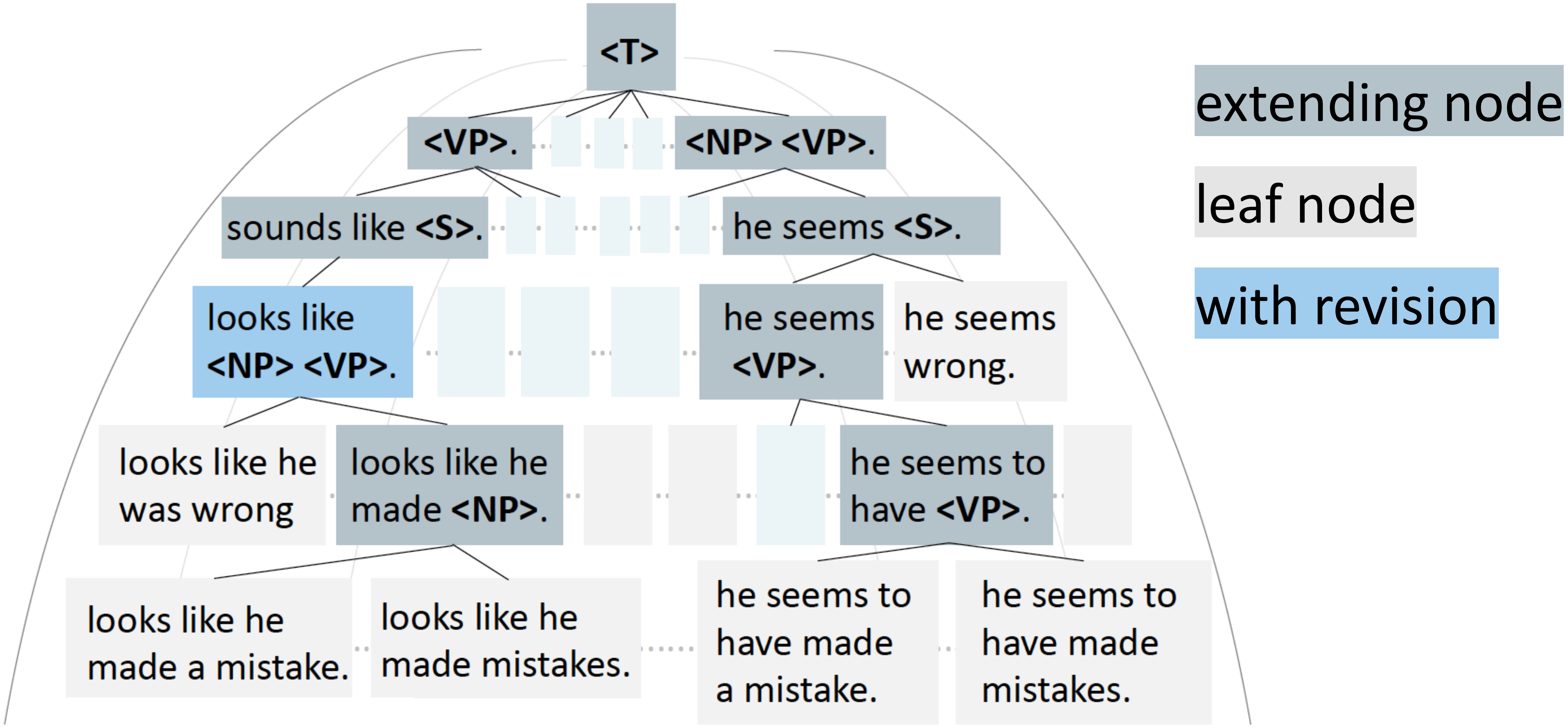


constituency **dependency** **w/o linguistic info**

[2] A Tree-based Decoder for Neural Machine Translation

Syntax-guided Generation

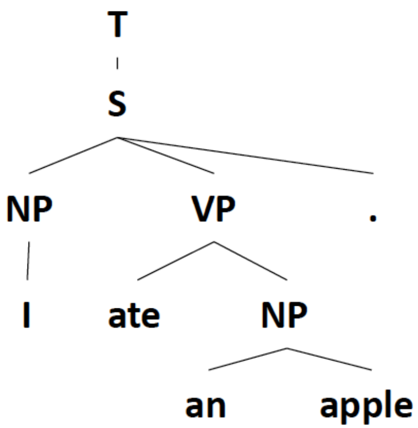
- Guide the decoding process using syntax information [3]



[3] Explicit Syntactic Guidance for Neural Text Generation

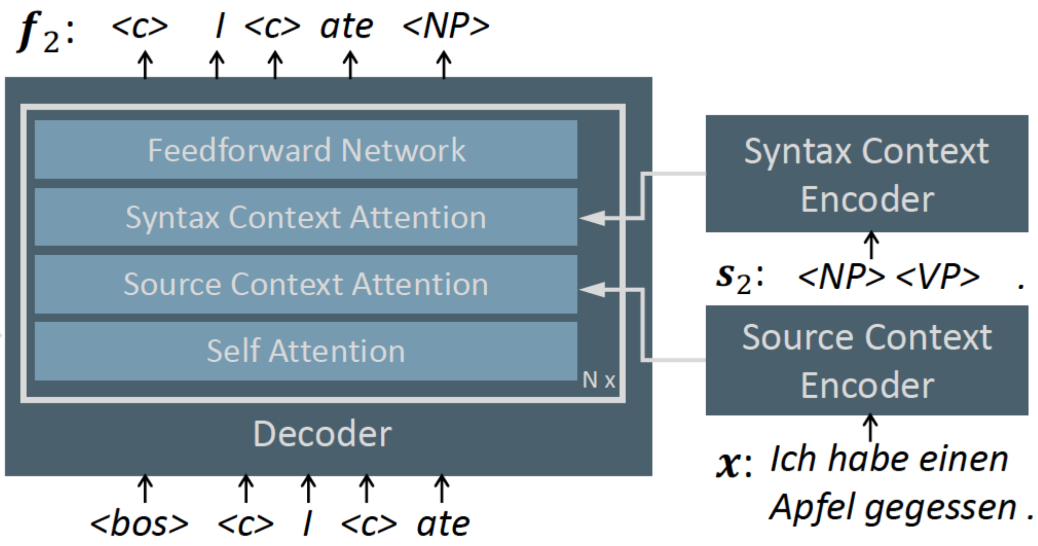
Method

Training data



d	$\mathbb{T}_d$	s_d	f_d
0	{(0,4,0,T)}	*<T>*	*<S>*
1	{(0,4,1,*S*)}	*<S>*	*<c> <NP> <VP> .*
2	{(0,0,2,*NP*), (1,3,2,*VP*)}	***<NP> <VP>*** *.*	*<c>* ***I*** *<c>* ***ate <NP>***
3	{(2,3,3,*NP*)}	*I ate <NP> .*	*<c> an apple*

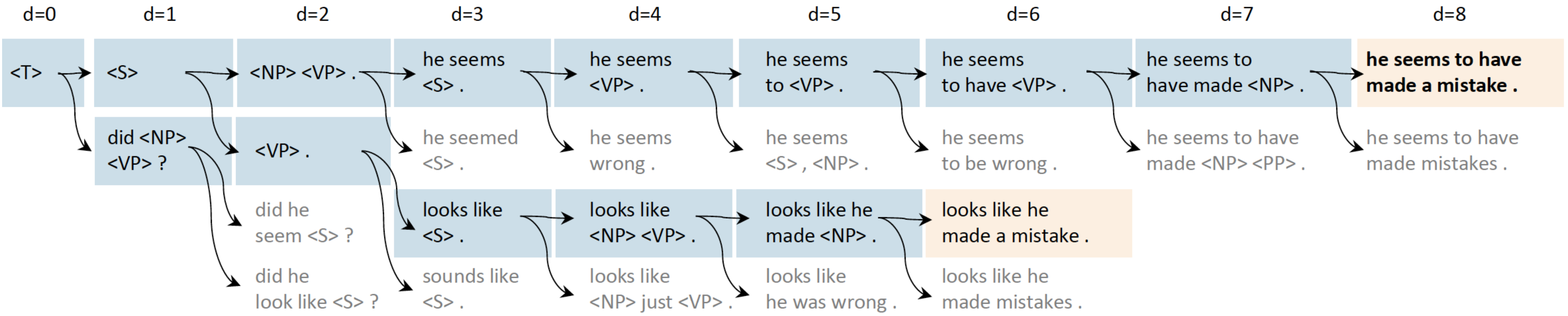
Decoder



[3] Explicit Syntactic Guidance for Neural Text Generation

Case Study

- Decoding process example [3]



[3] Explicit Syntactic Guidance for Neural Text Generation

Summary

- Leveraging the tree-structure information in the decoder/decoding is helps, especially in **low-resource scenarios**.
- The decoding process is also more **controllable/explainable**.

References

- [1] Towards String-to-Tree Neural Machine Translation
- [2] A Tree-based Decoder for Neural Machine Translation
- [3] Explicit Syntactic Guidance for Neural Text Generation

Linguistically Motivated Evaluation for Neural Machine Translation

Part of the EAMT 2024 Tutorial

Linguistically Motivated Neural Machine Translation

Haiyue Song

https://shyyhs.github.io

European Association for Machine Translation

National Institute of Information and Communications Technology

Evaluation

- Linguistic Evaluation Benchmark
- Linguistic Evaluation on the MT Output of GPT-4

Checklist

- [x] Ambiguity
- [x] Composition
- [ ] Punctuation
- [x] Verb tense

…

Linguistic Evaluation Benchmark

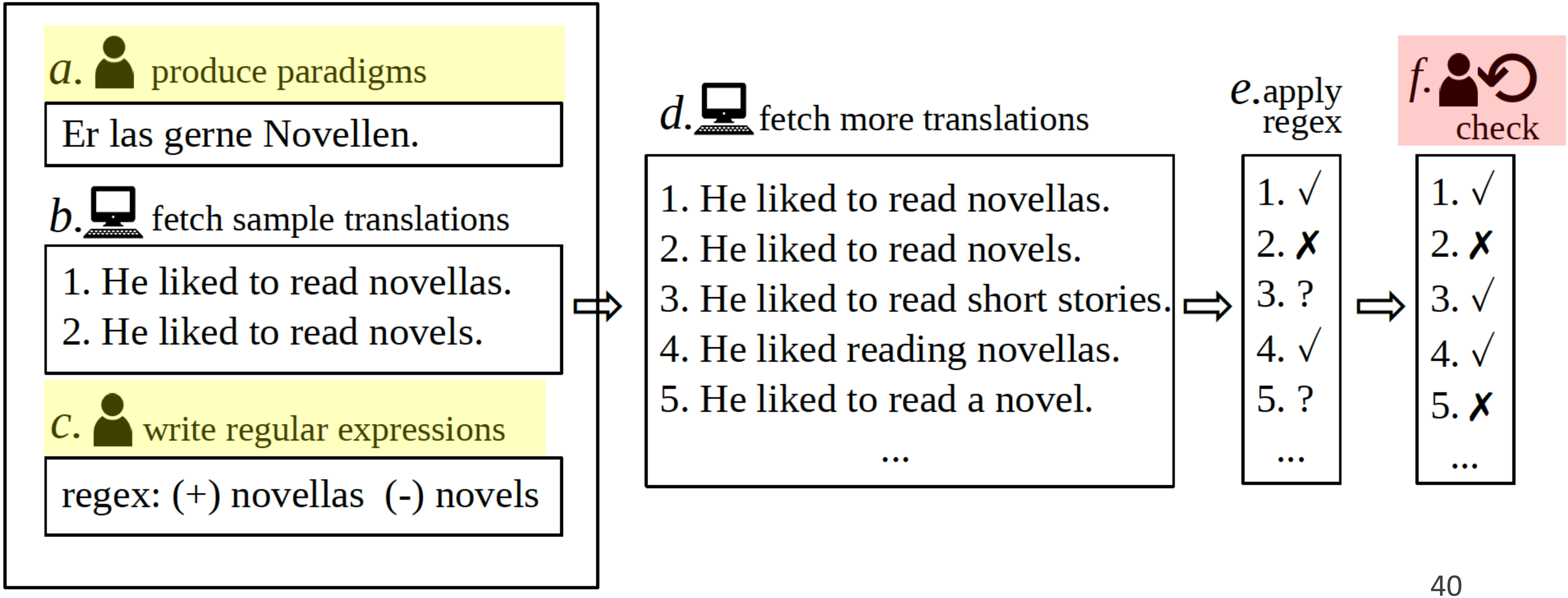
- Evaluate on **language-specific** linguistic phenomenon [1]

German→English	
Lexical Ambiguity	
Er las gerne Novellen.	
He liked to read novels.	fail
He liked to read novellas.	pass
Phrasal verb	
Warum starben die Dinosaurier aus?	
Why did the dinosaurs die?	fail
Why did the dinosaurs die out?	pass
Why did the dinosaurs become extinct?	pass
Ditransitive Perfect	
Ich habe Tim einen Kuchen gebacken.	
I have baked a cake.	fail
I baked Tim a cake.	pass

[1] Linguistically motivated Evaluation of the 2022 State-of-the-art Machine Translation Systems for three Language Directions

Linguistic Evaluation Benchmark

- A semi-automatic pipeline



Linguistic Evaluation on Metrics

- Check if the metric **favors the correct one** [2]

BERTScore

Good metric!

80 → He liked to read novels.

90 → He liked to read novellas.

XXXMetric

90 → Why did the dinosaurs die?

80 → Why did the dinosaurs die out?

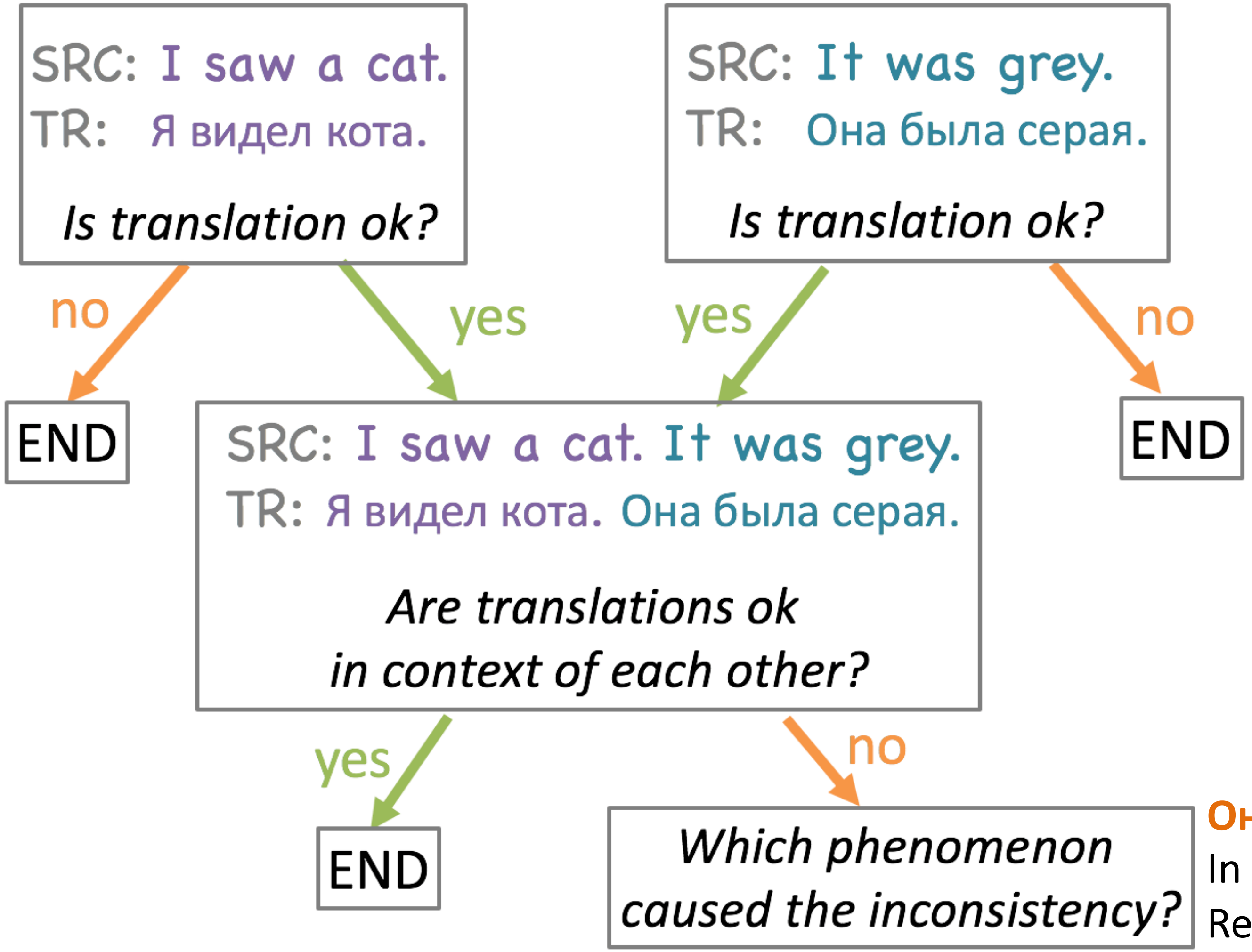
Lexical Ambiguity	
Er las gerne Novellen.	
He liked to read novels.	fail
He liked to read novellas.	pass
Phrasal verb	
Warum starben die Dinosaurier aus?	
Why did the dinosaurs die?	fail
Why did the dinosaurs die out?	pass

- High-performance metrics for En-De
  - BERTScore
  - COMET-22

[2] Linguistically Motivated Evaluation of Machine Translation Metrics based on a Challenge Set

Linguistic Evaluation on Context-Aware MT

- Translation should be contextual. [3]



Она and **Он** are indicative pronouns in Russian.
In the case of "cat", **Он** should be used.
Reference: **Он был серый.**

[3] When a Good Translation is Wrong in Context: Context-Aware Machine Translation Improves on Deixis, Ellipsis, and Lexical Cohesion

Phenomena in Context-Aware MT (1/3)

- Deixis
  - Referential expressions whose denotation depends on context.

EN Is someone putting you up to this? Are you being ... coerced?

RU Тебя кто-то подговорил? Вас принуждали?

Violation of T-V form consistency

- Informal form
- Formal form

[3] When a Good Translation is Wrong in Context: Context-Aware Machine Translation Improves on Deixis, Ellipsis, and Lexical Cohesion

Phenomena in Context-Aware MT (2/3)

- Ellipsis
  - The omission from a clause

Veronica, thank you, but you saw what happened. We all did.

Вероника, спасибо, но ты видела, что произошло. Мы все хотели.

“did” should be translated into a word meaning “saw” (видела) but wrongly into “want” (хотели)

[3] When a Good Translation is Wrong in Context: Context-Aware Machine Translation Improves on Deixis, Ellipsis, and Lexical Cohesion

Phenomena in Context-Aware MT (3/3)

- Lexical Cohesion
  - Named entity inconsistency

EN Not for Julia. Julia has a taste for taunting her victims.

RU Не для Джулии. Юлия умеет дразнить своих жертв.

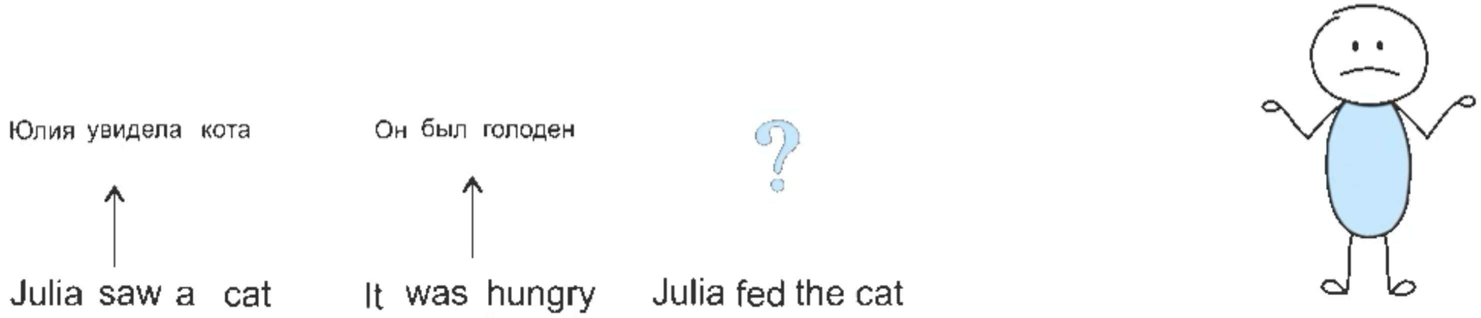
Translations of the name “Julia” are not consistent.

[3] When a Good Translation is Wrong in Context: Context-Aware Machine Translation Improves on Deixis, Ellipsis, and Lexical Cohesion

Method

- Better than simply concatenating contexts.

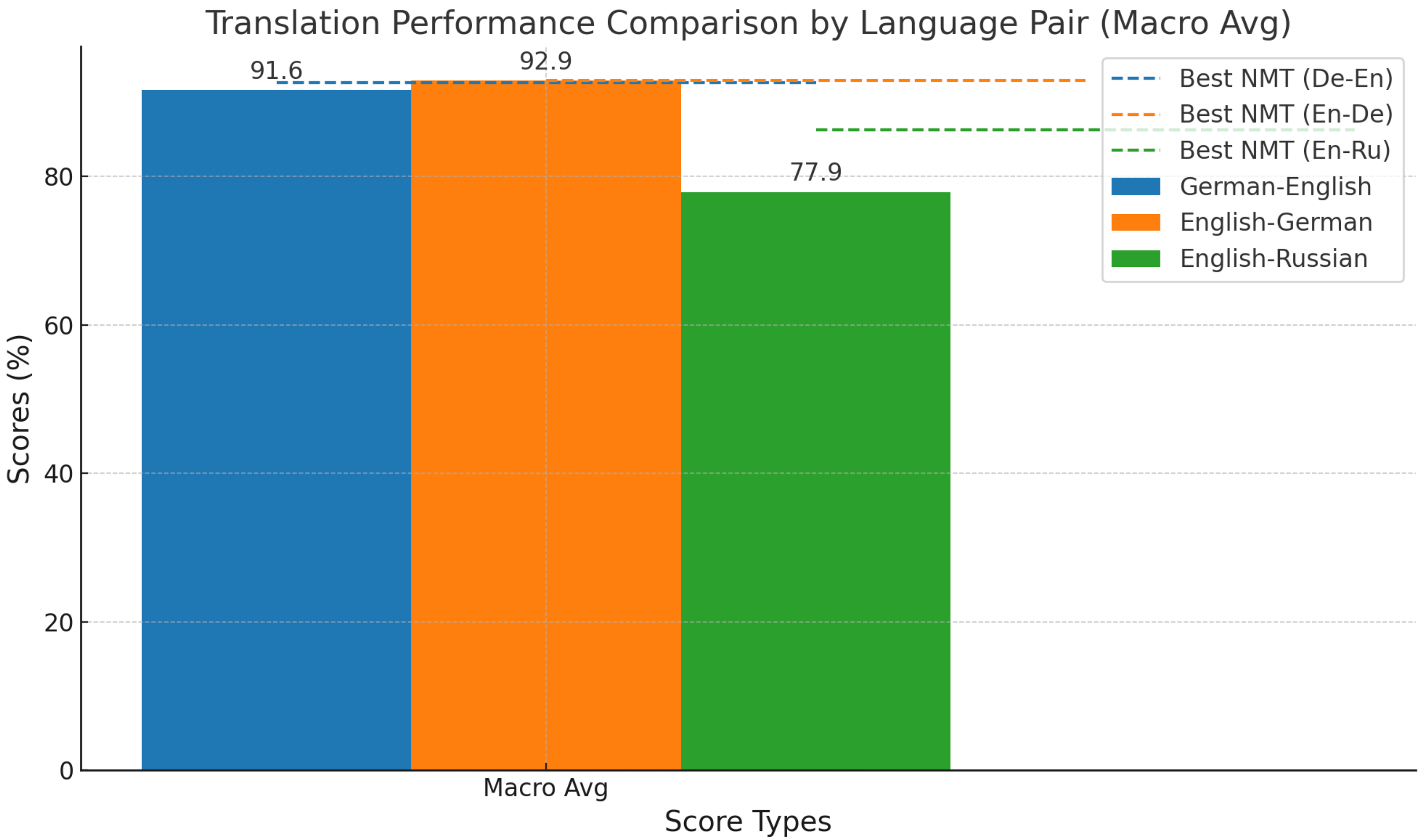
FIXME
Add results



[3] When a Good Translation is Wrong in Context: Context-Aware Machine Translation Improves on Deixis, Ellipsis, and Lexical Cohesion

Linguistic Evaluation on the MT Output of GPT-4

- Can GPT-4 outperforms traditional NMT models?
  - Comparable on high-resource directions
  - Not in lower-resource directions.



[4] Linguistically Motivated Evaluation of the 2023 State-of-the-art Machine Translation: Can GPT-4 Outperform NMT?

Summary

- The linguistic evaluation benchmark provides a more fine-grained evaluation of MT outputs.
  - However, it is still semi-automatic and requires human effort.
  - Better to add BERTScore/COMET-22 during evaluation which are consistent with this benchmark.
- Traditional MT systems are still better than GPT-4 especially in low-resource directions.

References

- [1] Linguistically motivated Evaluation of the 2022 State-of-the-art Machine Translation Systems for three Language Directions
- [2] Linguistically Motivated Evaluation of Machine Translation Metrics based on a Challenge Set
- [3] When a Good Translation is Wrong in Context: Context-Aware Machine Translation Improves on Deixis, Ellipsis, and Lexical Cohesion
- [4] Linguistically Motivated Evaluation of the 2023 State-of-the-art Machine Translation: Can GPT-4 Outperform NMT?