Data with Linguistic Knowledge for NMT

Part of the EAMT 2024 Tutorial

Linguistically Motivated Neural Machine Translation

Haiyue Song
https://shyyhs.github.io

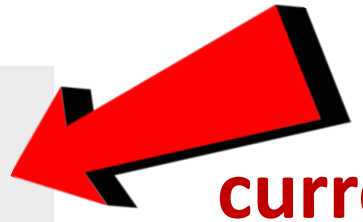
Roadmap

- Linguistic knowledge in

① **Model**

② **Data**

e.g. watch/ing
ab/normal/ly



current

③ **Decoding**

④ **Evaluation**

Inject Linguistic Knowledge into 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]

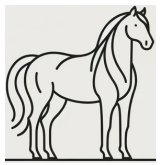
watch/ing sea/side

ab/normal/ly save/r/s

Character Decomposition [4]

Chinese 驰 → 马/也

run horse



Data from *Related Languages*

Script Mapping [5] [6]

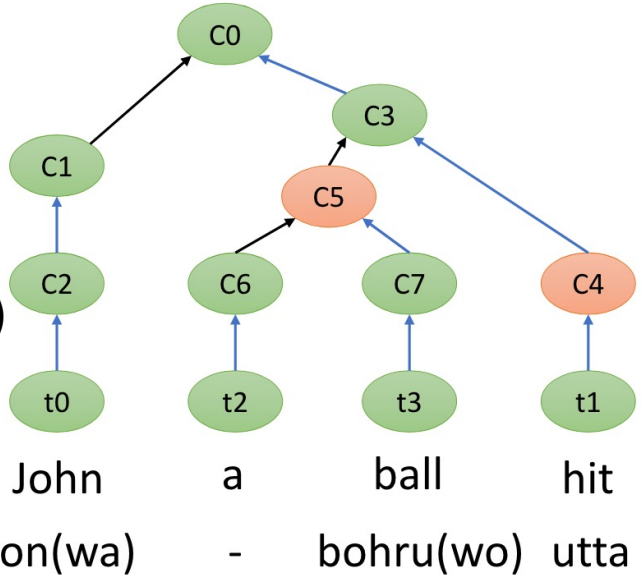
Chinese 她到塔皓湖去了

Romanization ta dao ta hao hu qu le

She went to Lake Tahoe.

Grammar Mapping [7]

Reorder English (SVO) to Japanese-like (SOV)



Word Segmentation: Motivation

- Add word boundary ⇨ **less ambiguity**

Japanese*

- 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 / 微信 / 、 / 支付宝 / 结合 / 的 / 趋势

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 comparison:
  - Juman++*

外国人参政権 → 外国 / 人 / 参政 / 権
right of foreigners to vote
foreign / people / suffrage / right

東京都知事 → 東京 / 都 / 知事
Tokyo governor
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

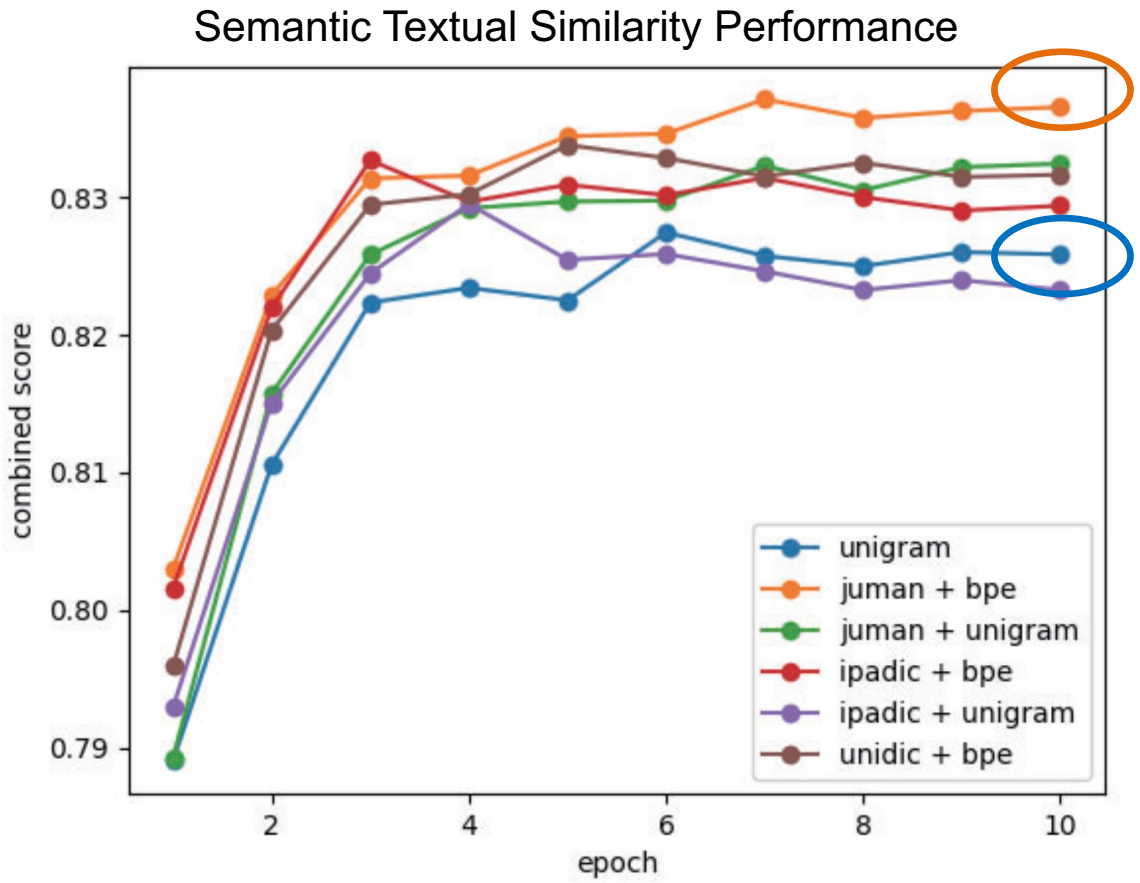
* Juman++: A Morphological Analysis Toolkit for Scriptio Continua
** SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing

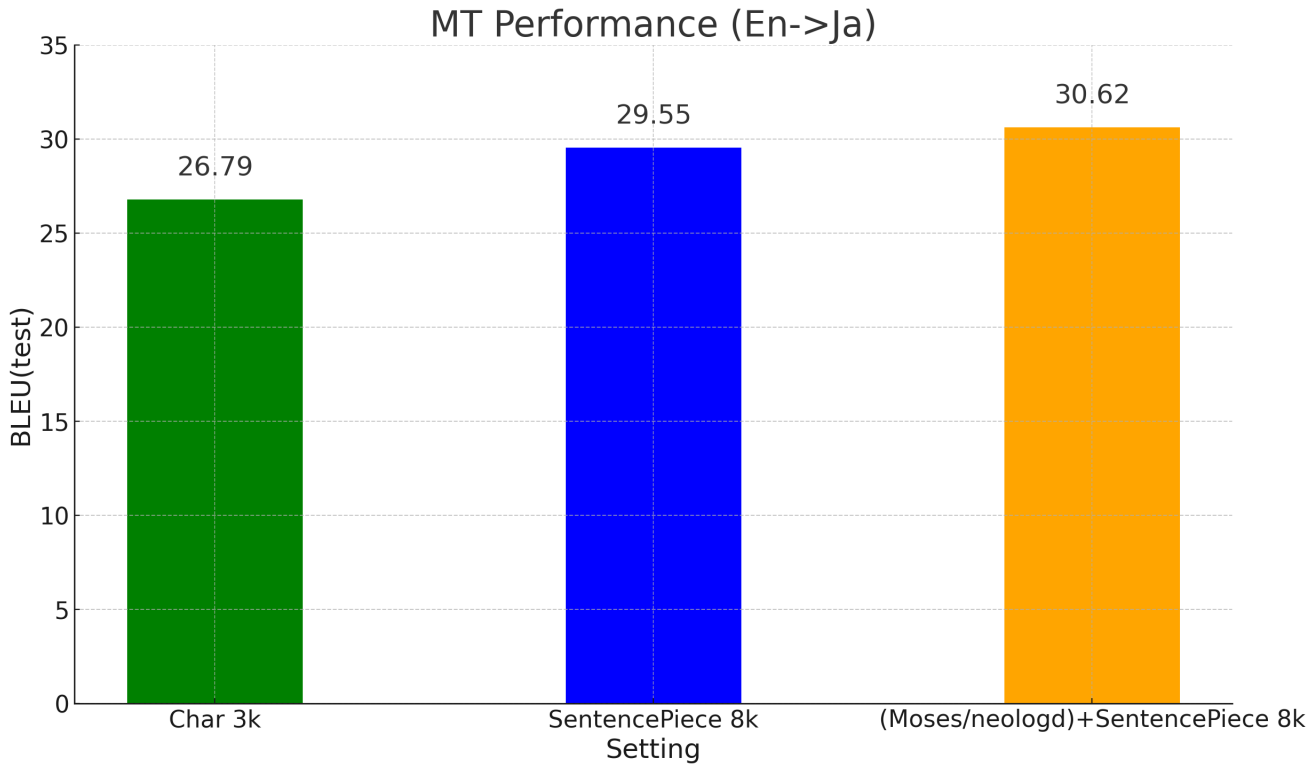
Word Segmentation: Improvements

KEY TAKEAWAY

- Downsteam task performance
  - **Juman + bpe/unigram** > **unigram**

For Chinese and Japanese, apply word segmenter before applying tokenization.





left: https://www.jstage.jst.go.jp/article/pjsai/JSAI2023/0/JSAI2023_2D6GS304/_pdf/-char/ja
right: https://github.com/google/sentencepiece/blob/master/doc/experiments.md

Conventional Subword Segmentation

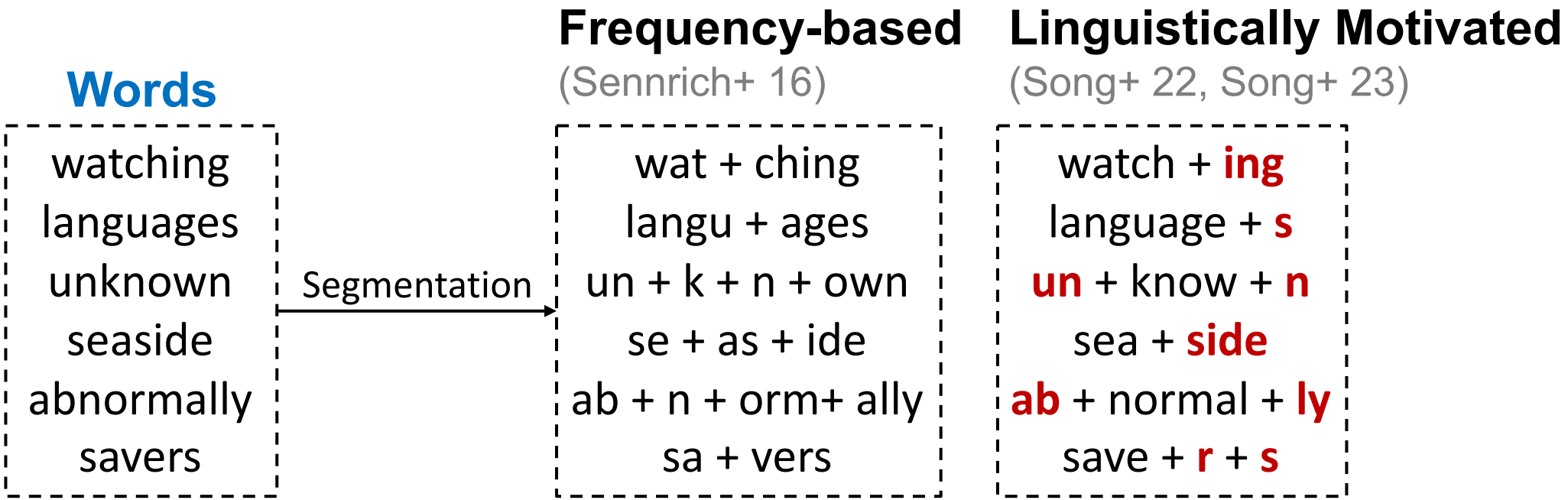
- NMT systems use **subwords** as the minimal unit.
- Compared to *word*, subwords handles **unseen words** by segmenting them into **seen subwords** in the subword vocabulary.

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*

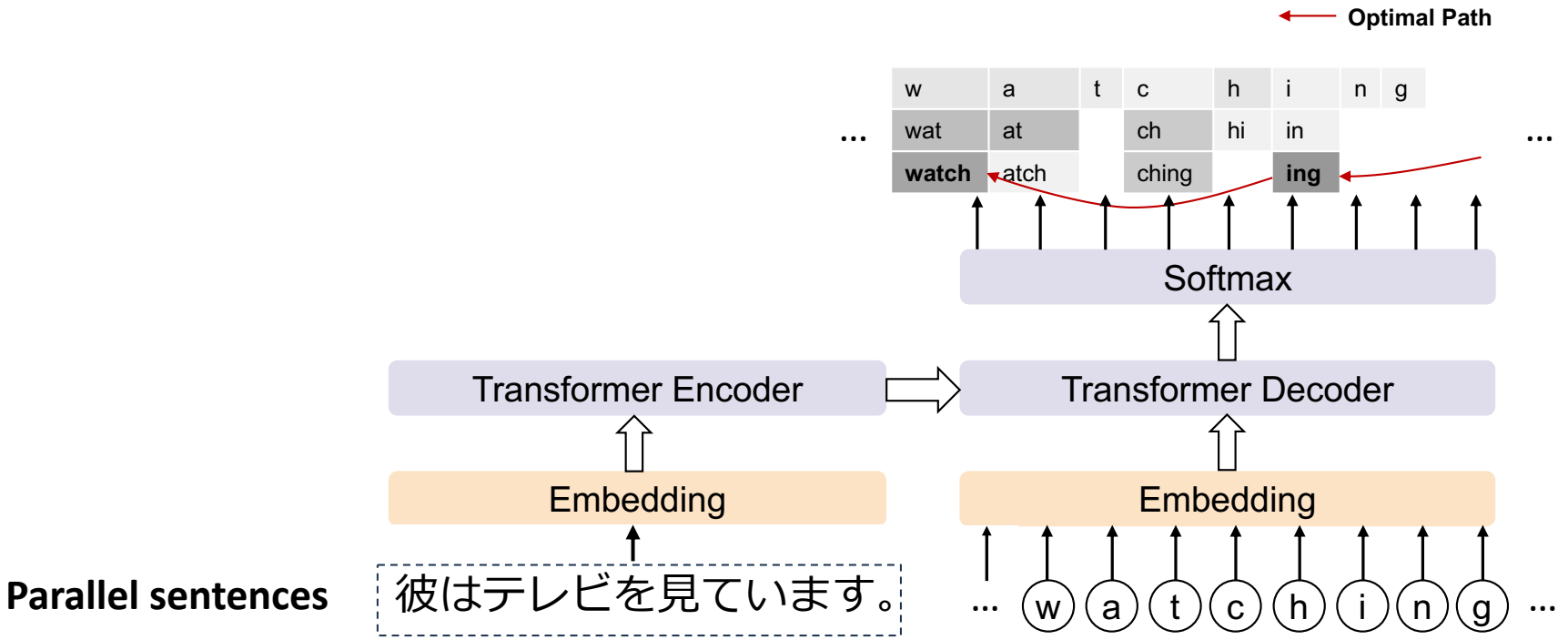
Linguistically Motivated Subword Segmentation

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



Subword Segmentation: Method (1/2)

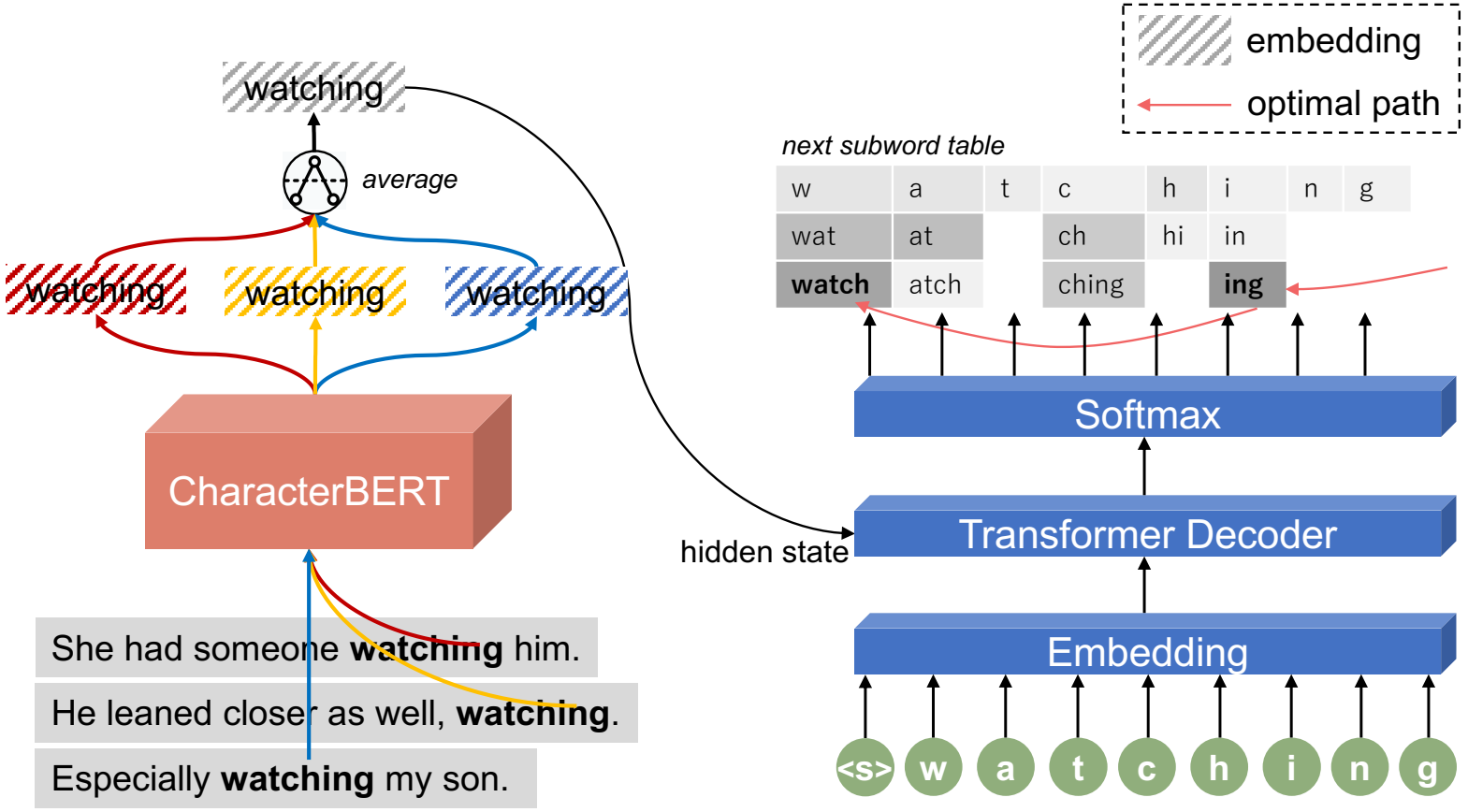
- Dynamic Programming Encoding (DPE)[2] is a neural segmenter trained on **parallel sentences**.
  - It maximizes the marginal likelihood of the **target sentence**.



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

Subword Segmentation: Method (2/2)

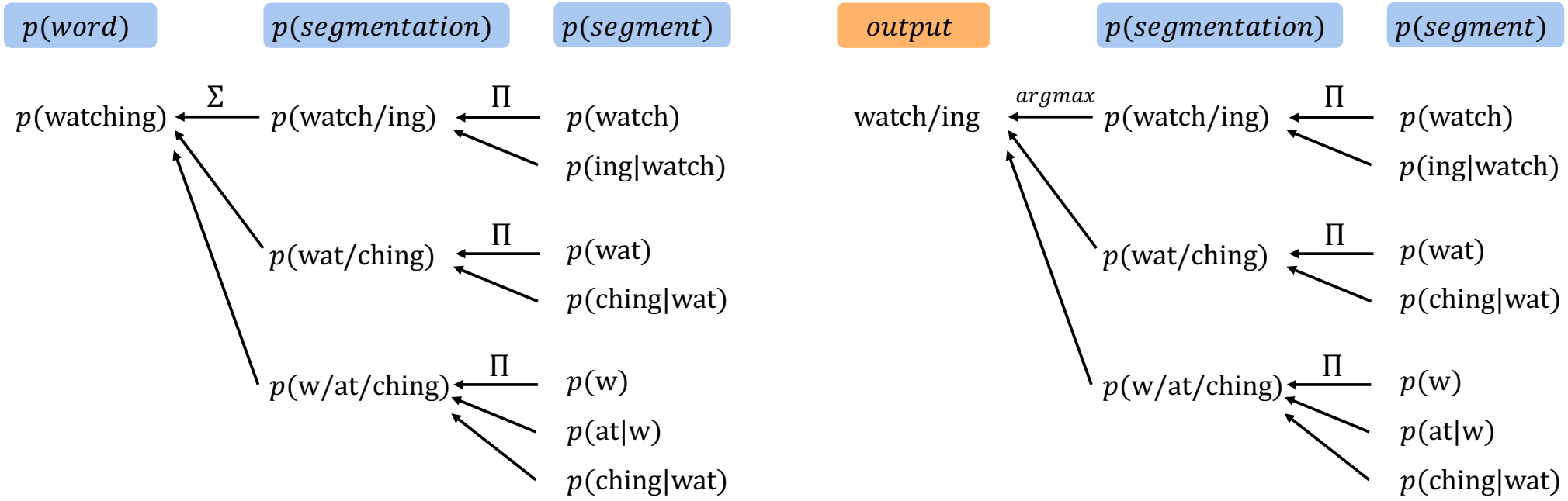
- BERTSeg
  - Uses semantic information from BERT embeddings.
  - It maximizes the marginal likelihood of the **target word**.



[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: Examples (1/2)

- BERTSeg vs. BPE

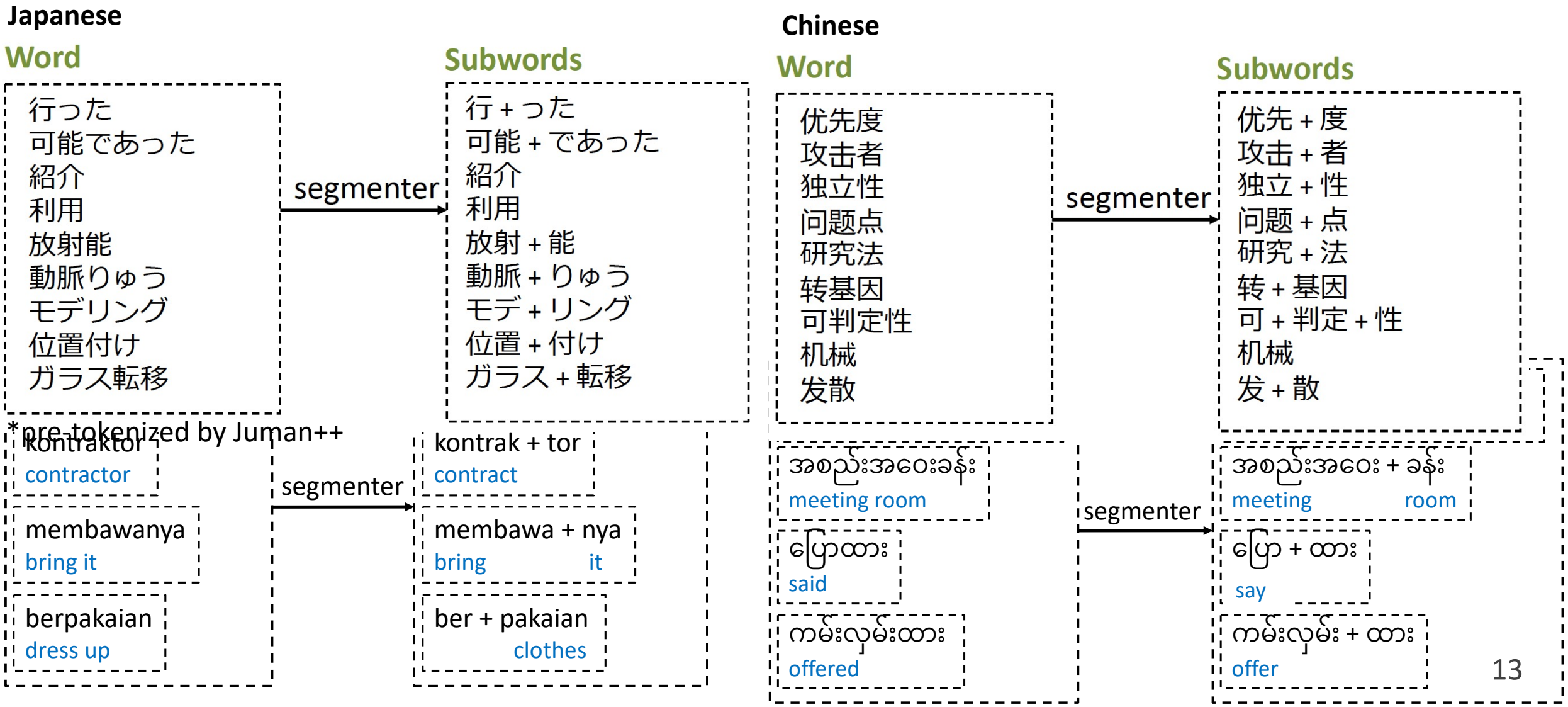
BERTSeg	BPE
Frequent words	
official/s	officials
edit/ion	edition
use/d	used
farm/er/s	far/mers
contribute/d	contrib/uted
normal/ly	norm/ally
seven/th	sevent/h
challenge/d	challeng/ed
over/night	o/vern/ight
language/s	langu/ages

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

Subword Segmentation: Examples (2/2)

- A **multilingual** segmenter using mBERT



Character Decomposition: Motivation

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

Character Decomposition

森 → 木 / 木 / 木

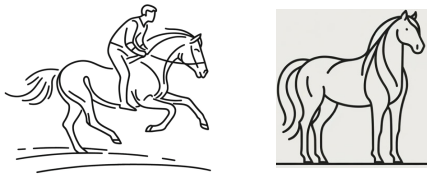
forest — tree — tree — tree

林 → 木 / 木

woods — tree — tree

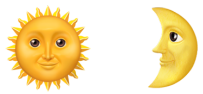
驰 → 马 / 也

run — horse



明 → 日 / 月

bright — sun — moon



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	丿乁一丨冂一丨ノ丶 丨冂一一丶一丨フ一丿ノ丶_1
	CN-character	风 景
Method in this paper	CN-ideograph	几乂 日亠口小_1
	CN-stroke	丿乁ノ丶 丨冂一一丶一丨フ一丿ノ丶_1
	EN	landscape

*Neural Machine Translation of Logographic Languages Using Sub-character Level Information

Character Decomposition: Results

- Best performance compared to word/character/stroke

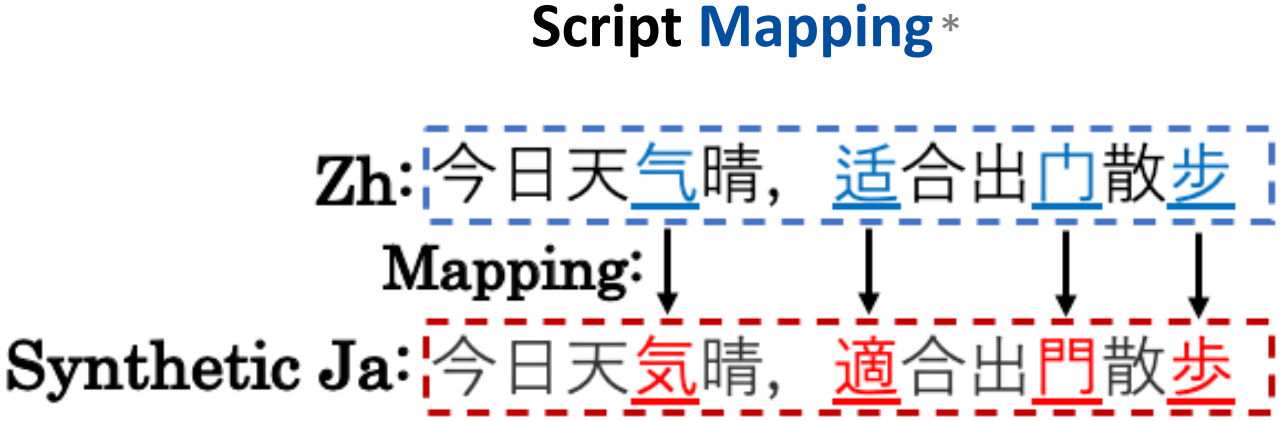
English-Chinese NMT		BLEU
EN_word	CN_word	11.8
EN_word	CN_character	10.3
EN_word	CN_ideograph	**14.6***
EN_word	CN_stroke	14.1*
Chinese-English NMT		**BLEU**
CN_word	EN_word	14.7
CN_character	EN_word	14.5
CN_ideograph	EN_word	**15.6***
CN_stroke	EN_word	15.5*

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
  - How to leverage data in **different scripts**?
  - How to leverage data in **different grammars**?

Script Mapping of Characters

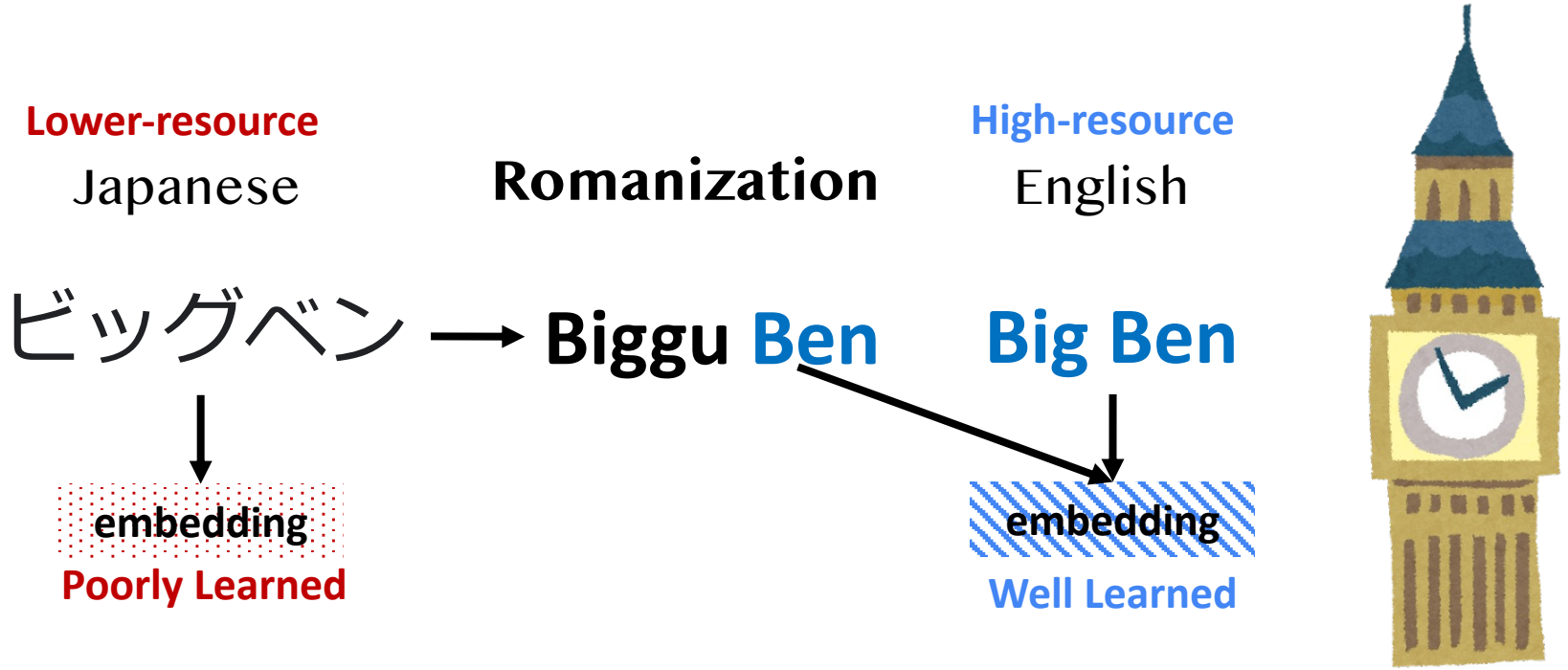
- From one language to another **related** language
  - By script mapping, where there is a *character-to-character dictionary*



*Pre-training via Leveraging Assisting Languages for Neural Machine Translation

Script Mapping by Romanization

- From non-Latin to Latin*



*On Romanization for Model Transfer Between Scripts in Neural Machine Translation

MT Performance using Romanization

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

		transfer from multilingual parent		
	base	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**

Grammar Mapping

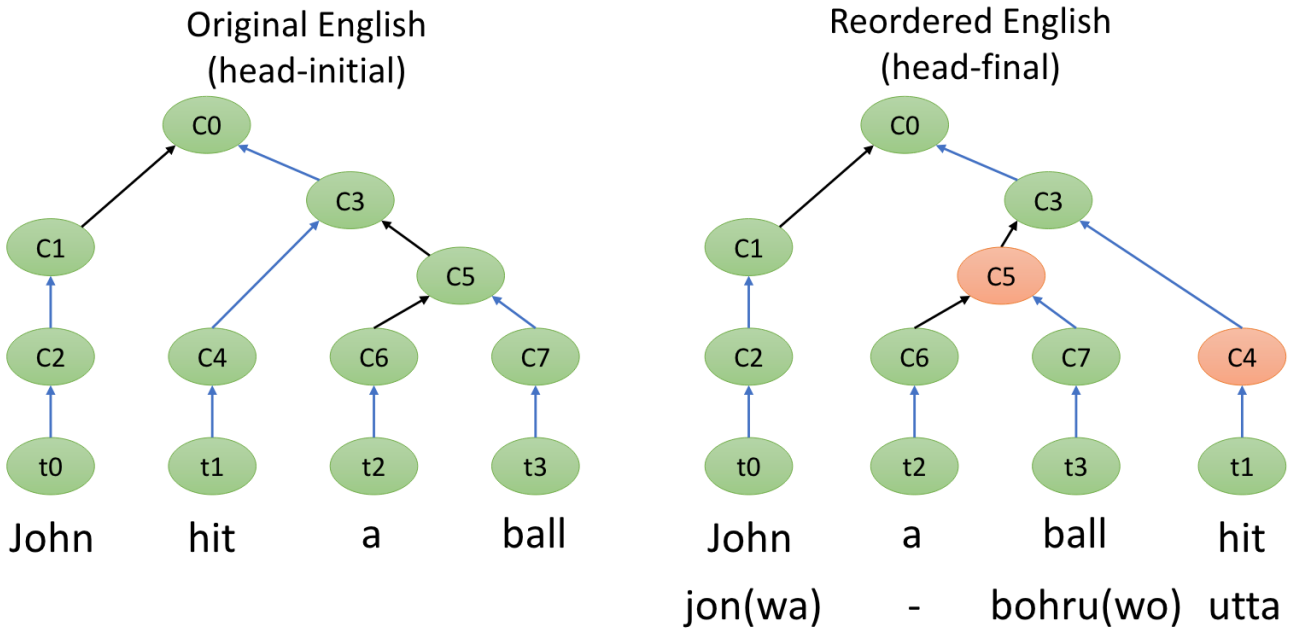
- English and Japanese have different word order
  - English: Subject–Verb–Object
  - Japanese: Subject–Object–Verb

John	**hit**	**a**	**ball**
John	a	ball	hit
jon(wa)	-	bohru(wo)	utta

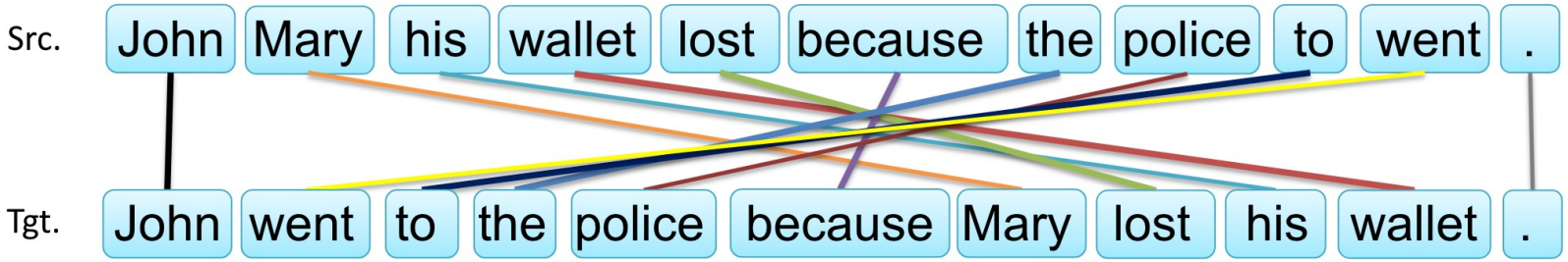
- Motivation
  - map English sentence to “English sentence with Japanese SOV word order
  - learn English with SOV order->English to simulate Japanese->English

Grammar Mapping

- Generate the reordered English data from head-initial to head-final



- Use it in **pre-training** improves the Japanese->English performance



Summary

- Linguistic knowledge can be injected into 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
  - Grammar mapping

Open Questions

- Character-level/Byte level tokenization
  - Character-level contains all information in theory
  - But it underperforms subword based methods
  - 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 is hard to leverage that knowledge *during generation*
  - Romanization hurts the performance of the original language
  - How to transfer between languages with **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
[7] Linguistically-driven Multi-task Pre-training for Low-resource Neural Machine Translation

Linguistically Aware Decoding

Part of the EAMT 2024 Tutorial

Linguistically Motivated Neural Machine Translation

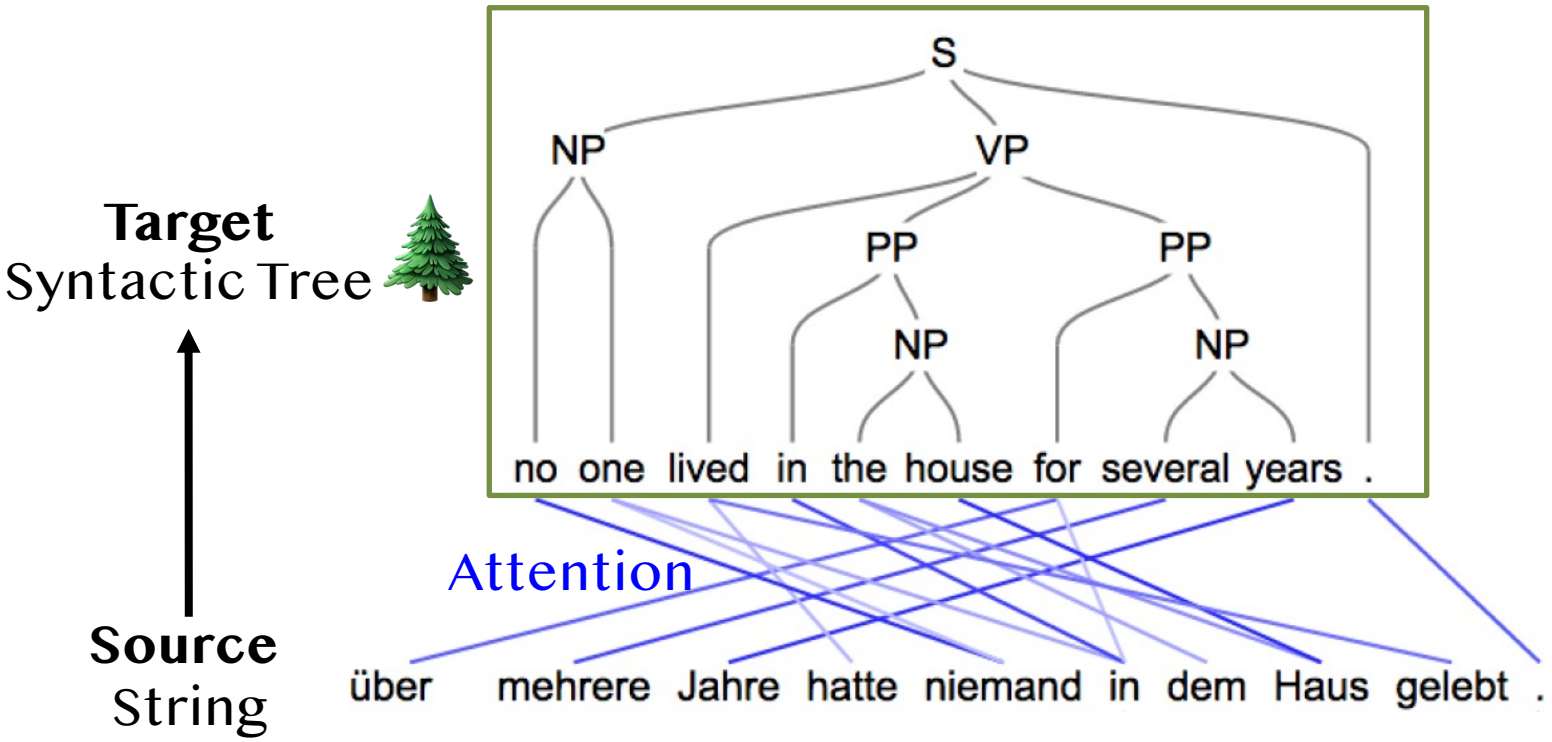
Haiyue Song
https://shyyhs.github.io

Decoding

- String-to-Tree Decoding
- Structural Template Prediction

String to Syntactic Tree: Overview

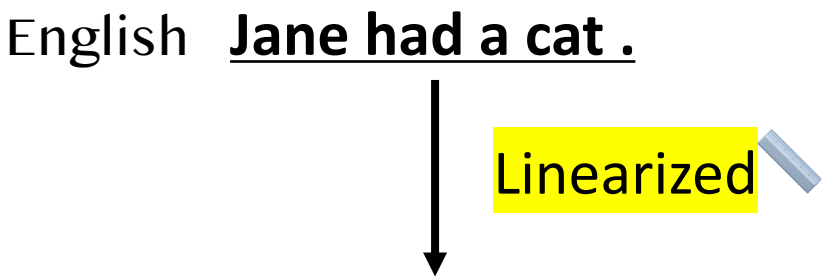
- Incorporate syntactic tree information [1]



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

String to Syntactic Tree: Method

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



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

String to Syntactic Tree: Results

- Helpful in low-resource scenarios
- No large improvement in high-resource scenarios

166k parallel sentences low-resource scenario

German➡️English

system	newstest2015	newstest2016
bpe2bpe	13.81	14.16
bpe2tree	14.55 ↑	16.13 ↑↑

4.5M parallel sentences

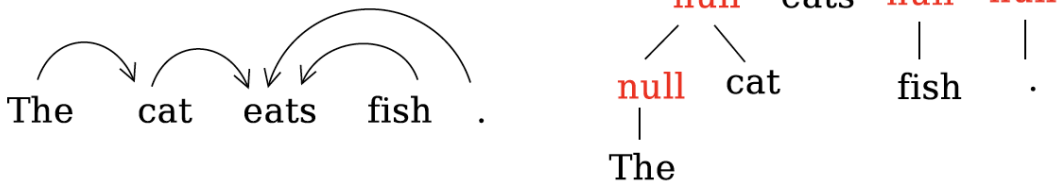
German➡️English

system	newstest2015	newstest2016
bpe2bpe	27.33	31.19
bpe2tree	27.36 ~	32.13 ↑

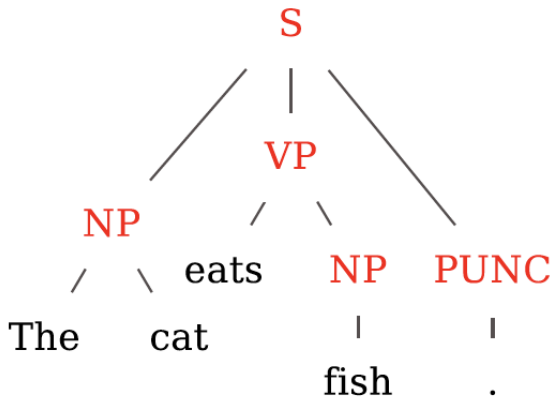
String to Any Tree: Overview

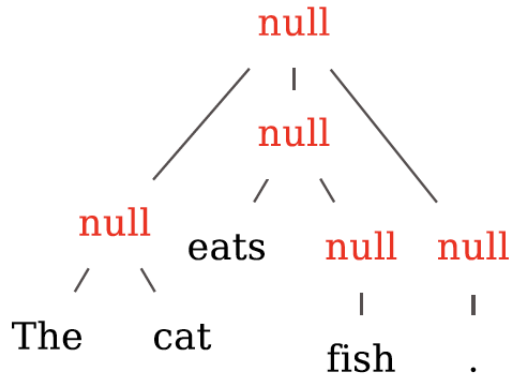
- Incorporate any tree structure information [2]

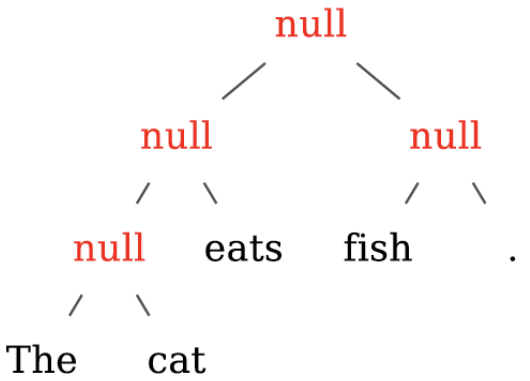
Example of one parser

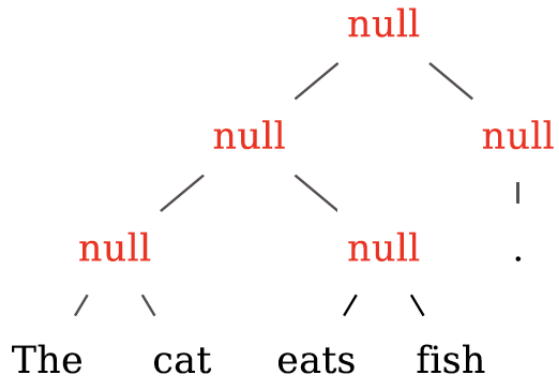


Different parsers, or even not from parser









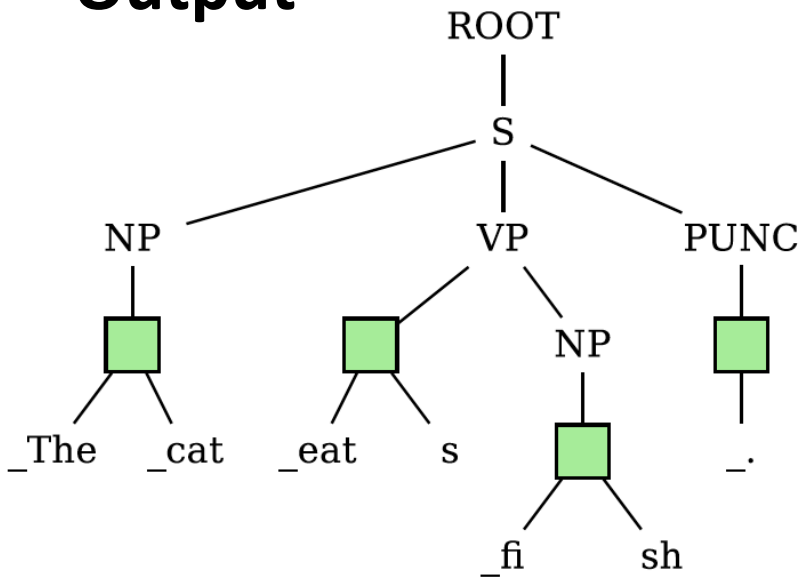
Balanced Binary Tree

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

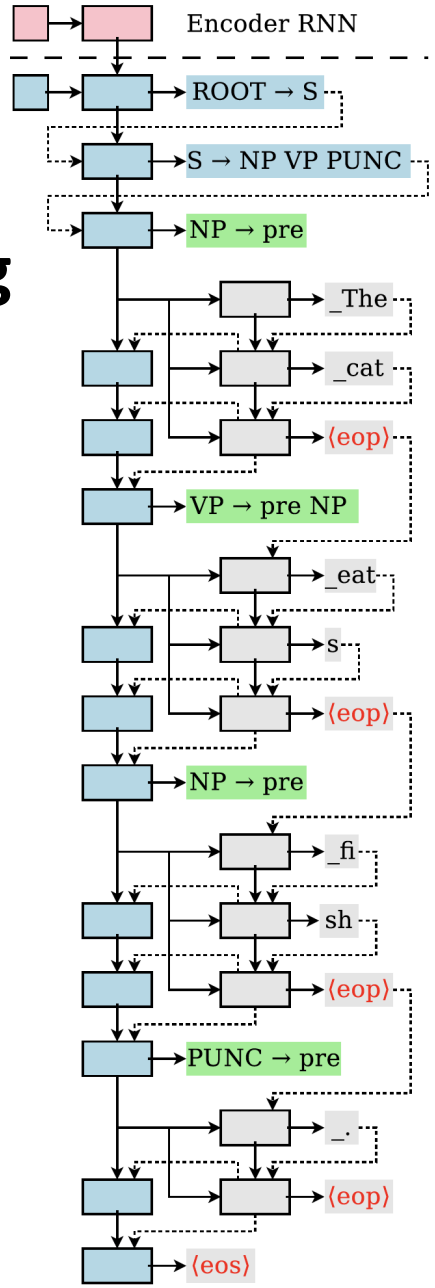
String to Any Tree: Method

- Tree-structure-aware decoder [2]

Output



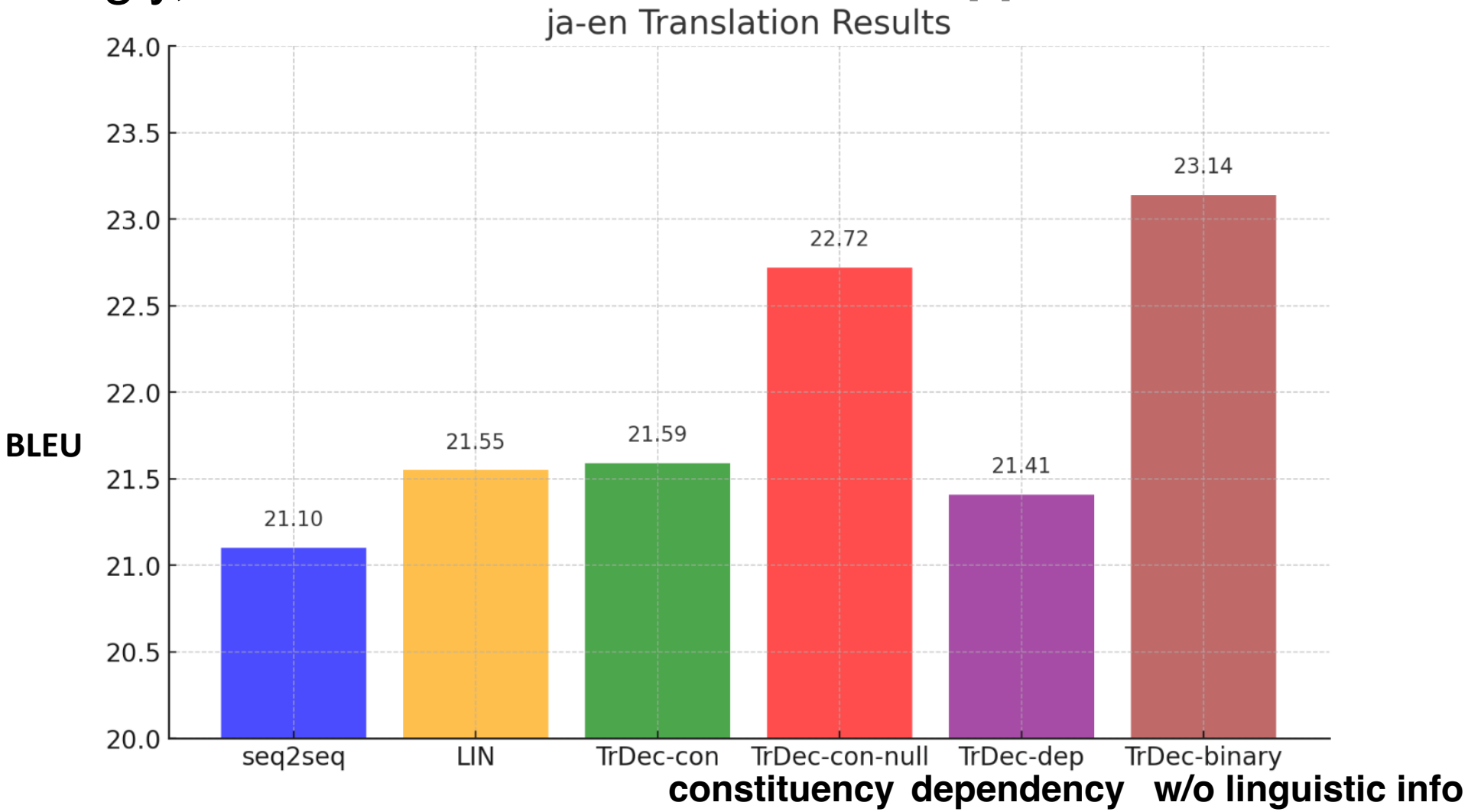
Decoding process



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

String to Any Tree: Results

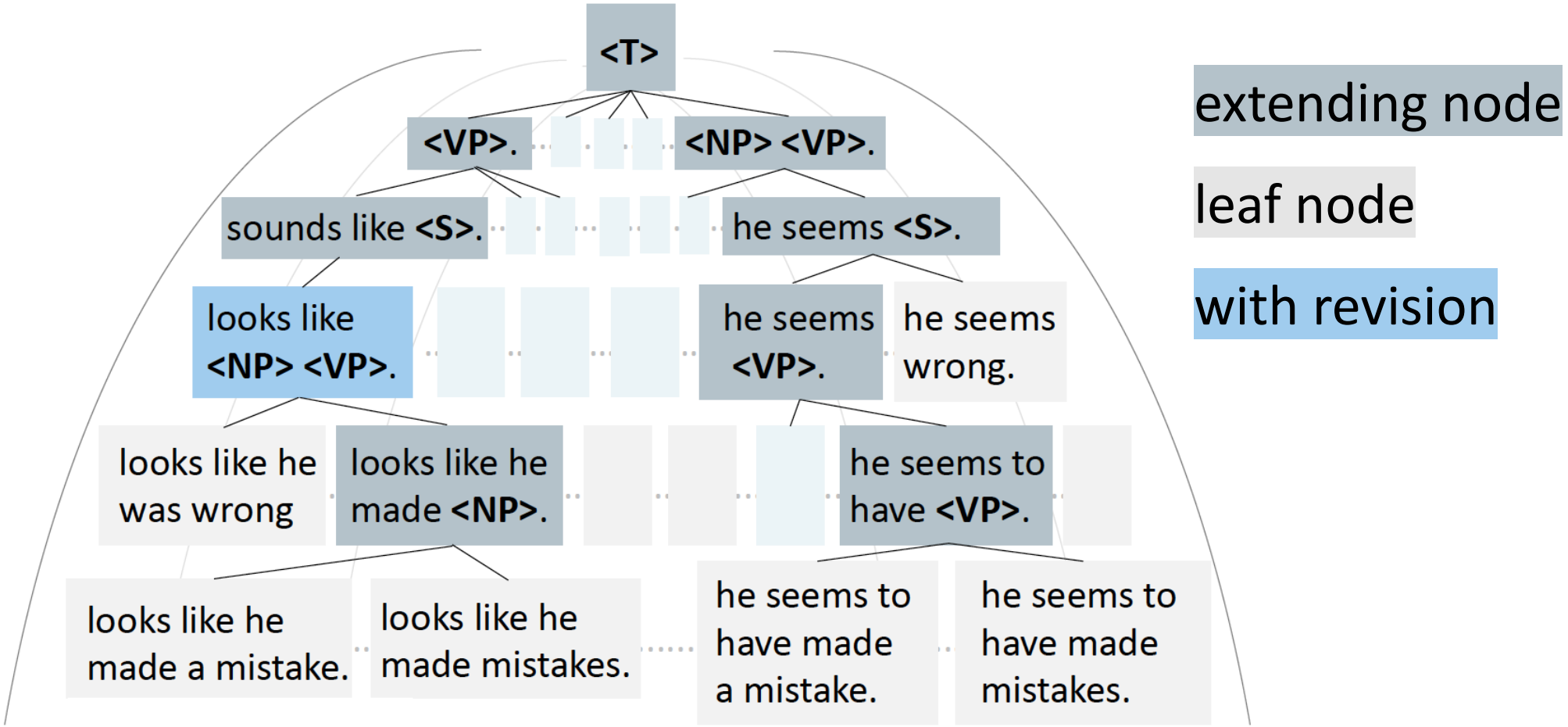
- Surprisingly, balanced tree also works well [2]



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

Syntax-guided Generation: Overview

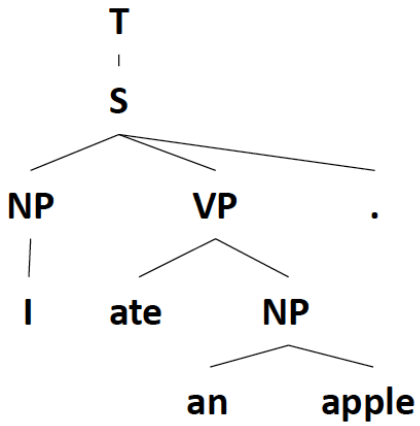
- Guide the decoding process using syntax information [3]



[3] Explicit Syntactic Guidance for Neural Text Generation

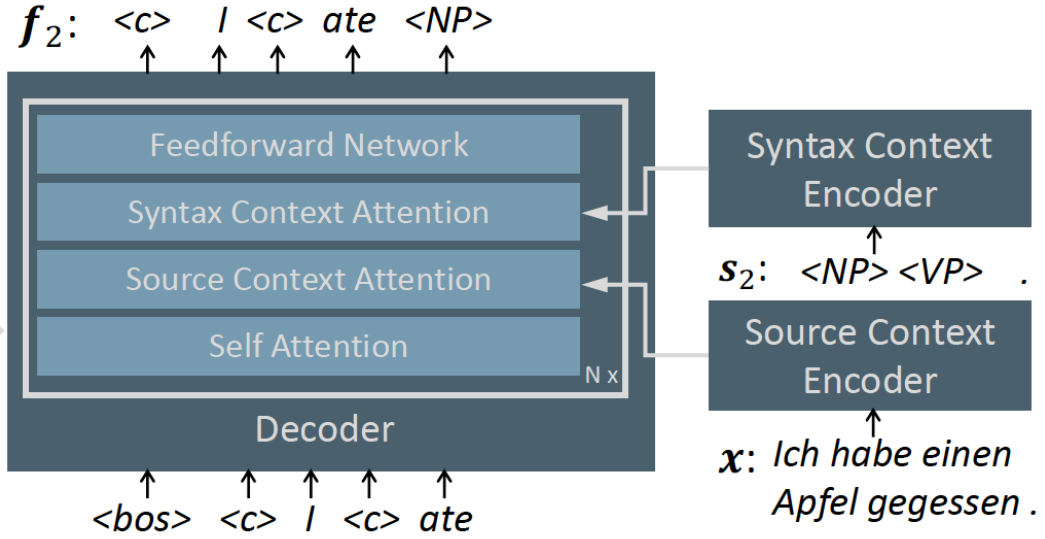
Syntax-guided Generation: Method

Training data



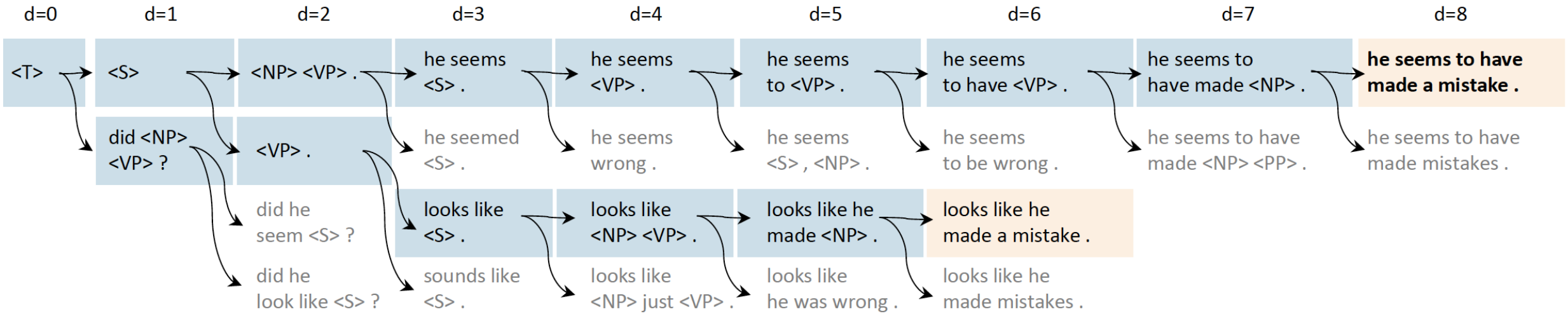
d	$\mathbb{T}_d$	s_d	f_d
0	{(0,4,0,T)}	<T>	<S>
1	{(0,4,1,S)}	<S>	<c> <NP> <VP> .
2	{(0,0,2,NP), (1,3,2,VP)}	<NP> <VP> .	<c> I <c> ate <NP>
3	{(2,3,3,NP)}	I ate <NP> .	<c> an apple

Decoder



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

Evaluation

- Linguistic Evaluation Benchmark
  - Construction
  - Evaluation on MT output of GPT-4
  - Evaluation on *MT Metrics*
- Context-aware MT evaluation

Linguistic Evaluation Benchmark

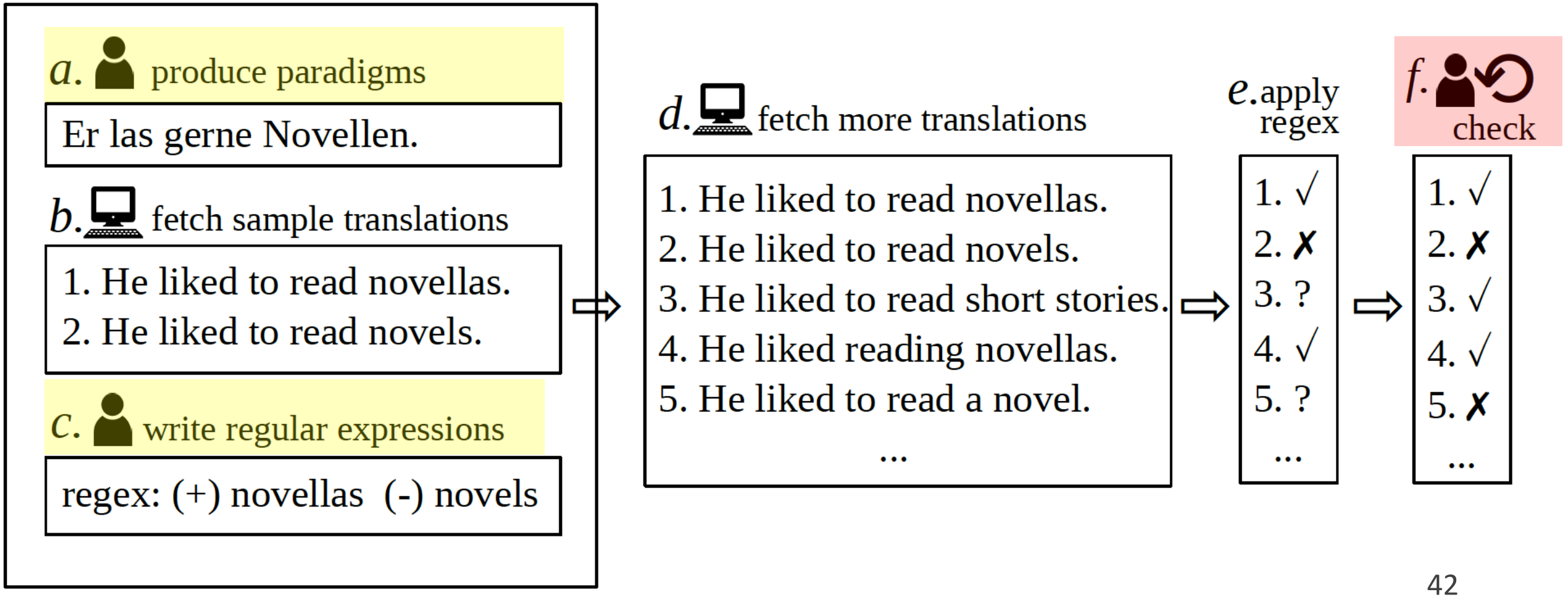
- Evaluate on **language-specific** linguistic phenomenon [1]

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: Construction

- A **semi-automatic** pipeline



Linguistic Evaluation Benchmark: Category

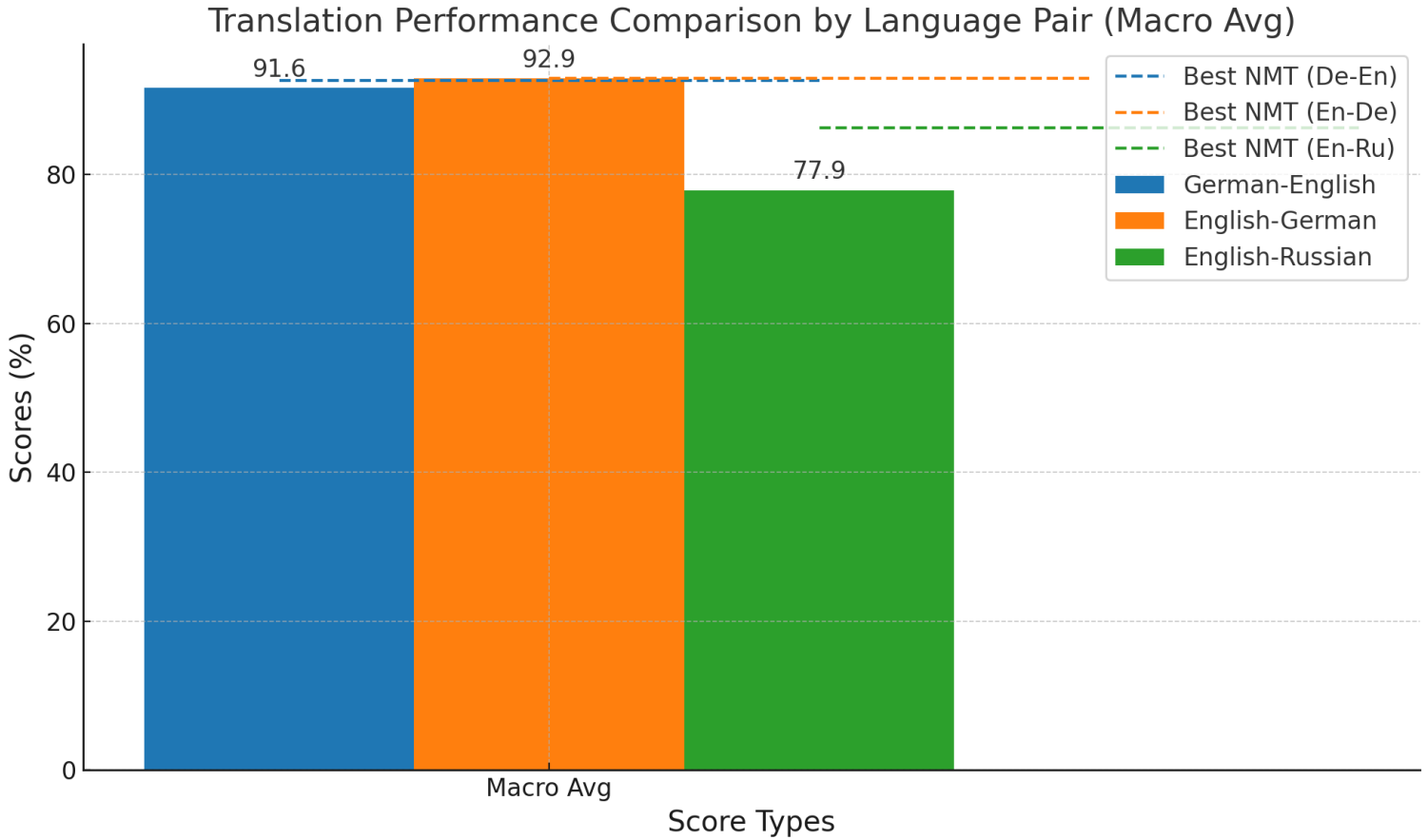
For De-En,
14 categories,
106 phenomena,
and >5,000 sentences.

Test set	Test sentences	Categories	Phenomena
De–En	∼5,500	14	106
En–De	∼4,400	13	110
En–Ru	∼300	12	51

categ	count
Ambiguity	81
Lexical ambiguity	63
Structural ambiguity	18
Composition	49
Compound	29
Phrasal verb	20
Coordination & ellipsis	56
Gapping	19
Right node raising	19
Sluicing	18
False friends	36
Function word	69
Focus particle	24
Modal particle	25
Question tag	20
LDD & interrogatives	149
Extended adjective construction	14
Extraposition	17
Multiple connectors	19

Linguistic Evaluation on the MT Output of GPT-4

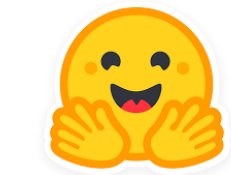
- Can GPT-4 outperforms traditional NMT models?
  - Comparable on high-resource directions
  - Not in lower-resource directions.



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

Linguistic Evaluation on Metrics

- Which MT quality estimation **Metrics** is better?
- Check if the metric **favors the correct one** [2]



BERTScore

Good metric!

80

90

XXXMetric

90

80

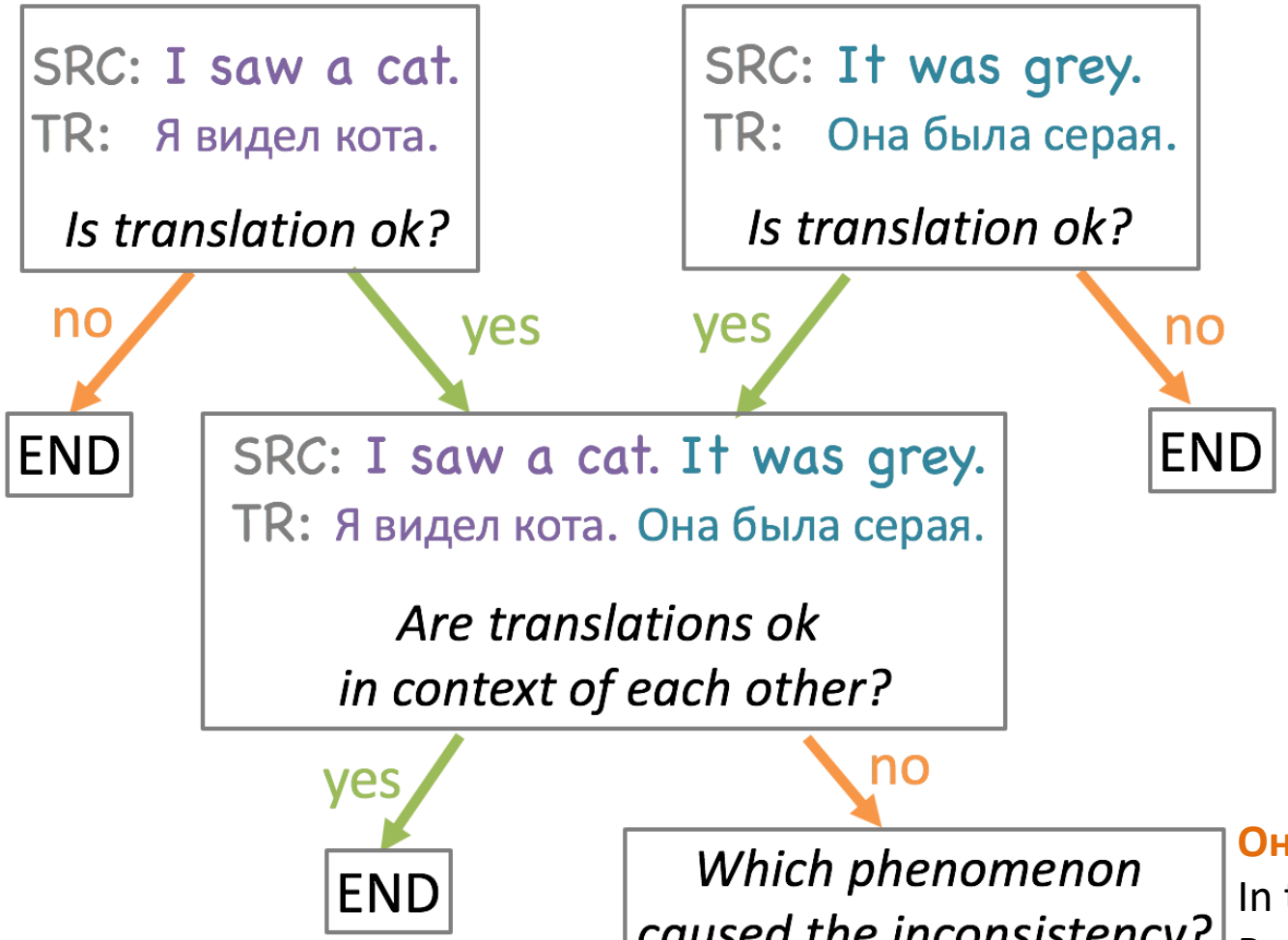
Lexical Ambiguity	
Er las gerne Novellen.	
He liked to read novels.	fail
He liked to read novellas.	pass
Phrasal verb	
Warum starben die Dinosaurier aus?	
Why did the dinosaurs die?	fail
Why did the dinosaurs die out?	pass

- 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]
  - Possible that translation for a single sentence is **correct** but the combined translation is **wrong**.



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 принуж-дали?

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.

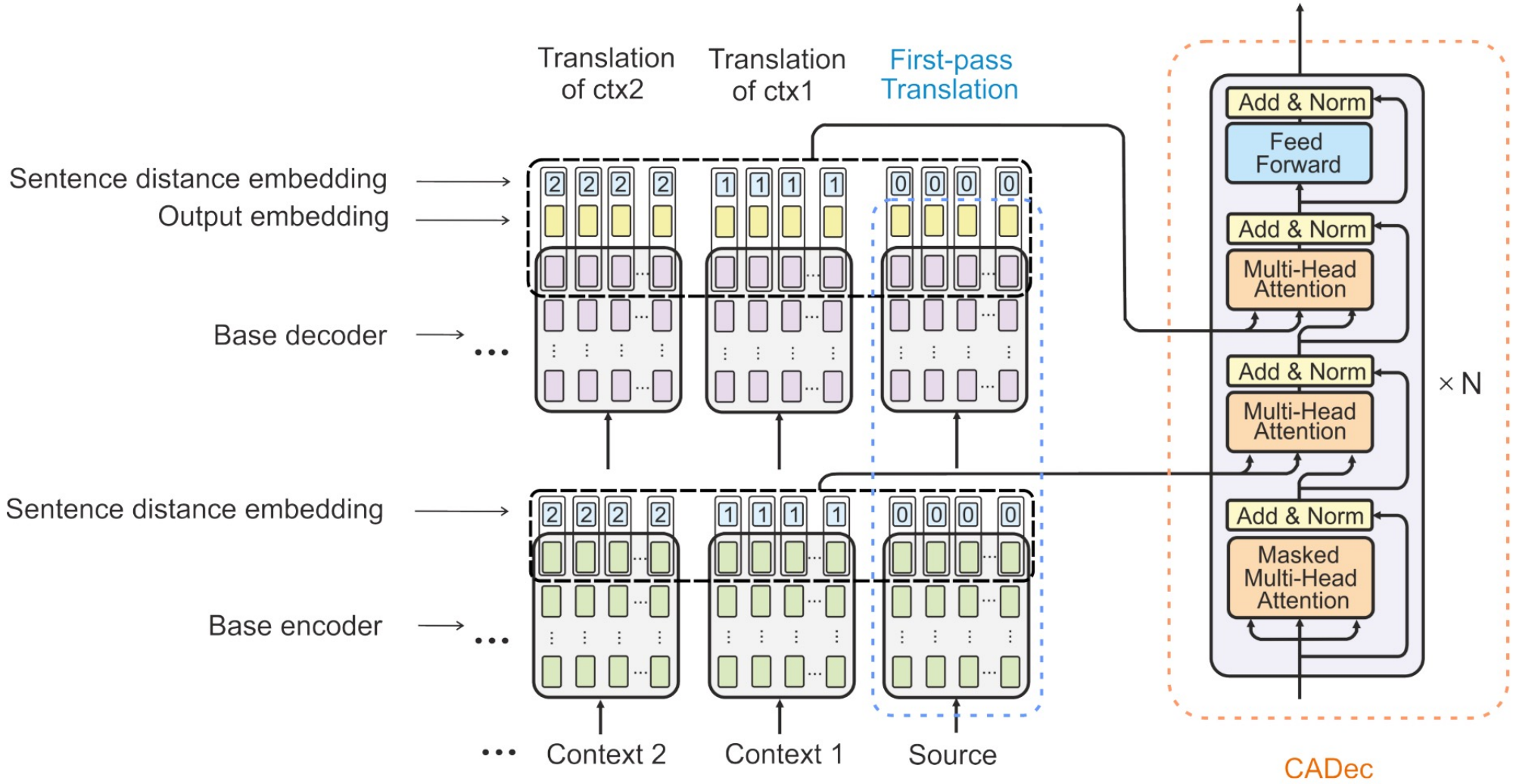
RU Не для Джулии. Юлия умеет дразнить своих жертв.

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

Context-Aware MT: Method

- Feed context/first-pass translations with distance embeddings



Context-Aware MT: Results

- BLEU is comparable but linguistically evaluation results are good

model	BLEU
baseline (1.5m)	29.10
baseline (6m)	**32.40**
concat	31.56
s-hier-to-2.tied	26.68
CADec	**32.38**

		latest relevant context		
	total	1st	2nd	3rd
deixis				
baseline	50.0	50.0	50.0	50.0
concat	**83.5**	**88.8**	**85.6**	**76.4**
s-hier-to-2.tied	60.9	83.0	50.1	50.0
CADec	81.6	84.6	84.4	75.9
lexical cohesion				
baseline	45.9	46.1	45.9	45.4
concat	47.5	48.6	46.7	46.7
s-hier-to-2.tied	48.9	53.0	46.1	45.4
CADec	**58.1**	**63.2**	**52.0**	**56.7**

Table 7: Accuracy for deixis and lexical cohesion.

	ellipsis (infl.)	ellipsis (VP)
baseline	53.0	28.4
concat	**76.2**	76.6
s-hier-to-2.tied	66.4	65.6
CADec	72.2	**80.0**

Table 8: Accuracy on ellipsis test set.

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 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?