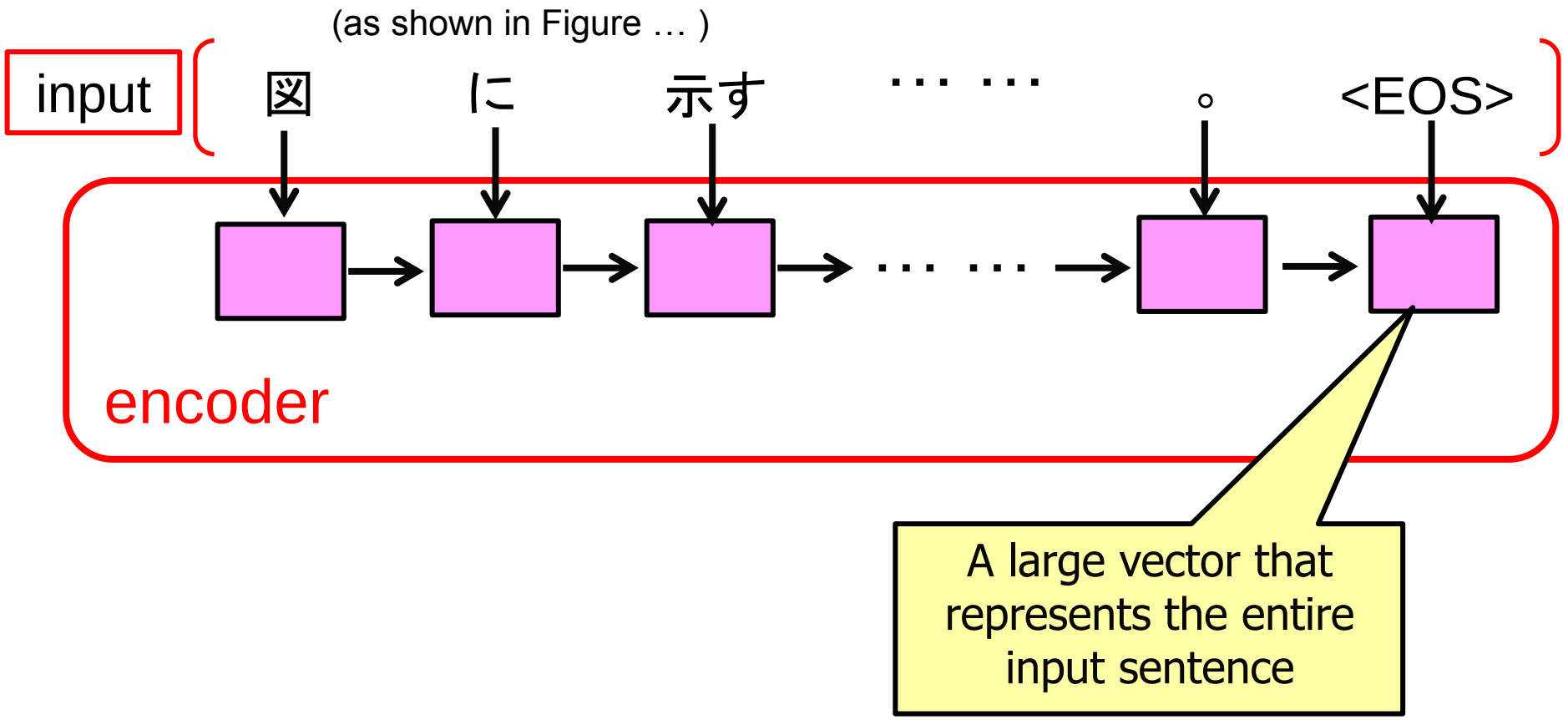
Translation of Patent Sentences with a Large Vocabulary of Technical Terms Using Neural Machine Translation

Zi Long, Takehito Utsuro
University of Tsukuba
Tomoharu Mitsuhashi
Japan Patent Information Organization
Mikio Yamamoto
University of Tsukuba

WAT2016, December 12, 2016 @ Osaka, Japan

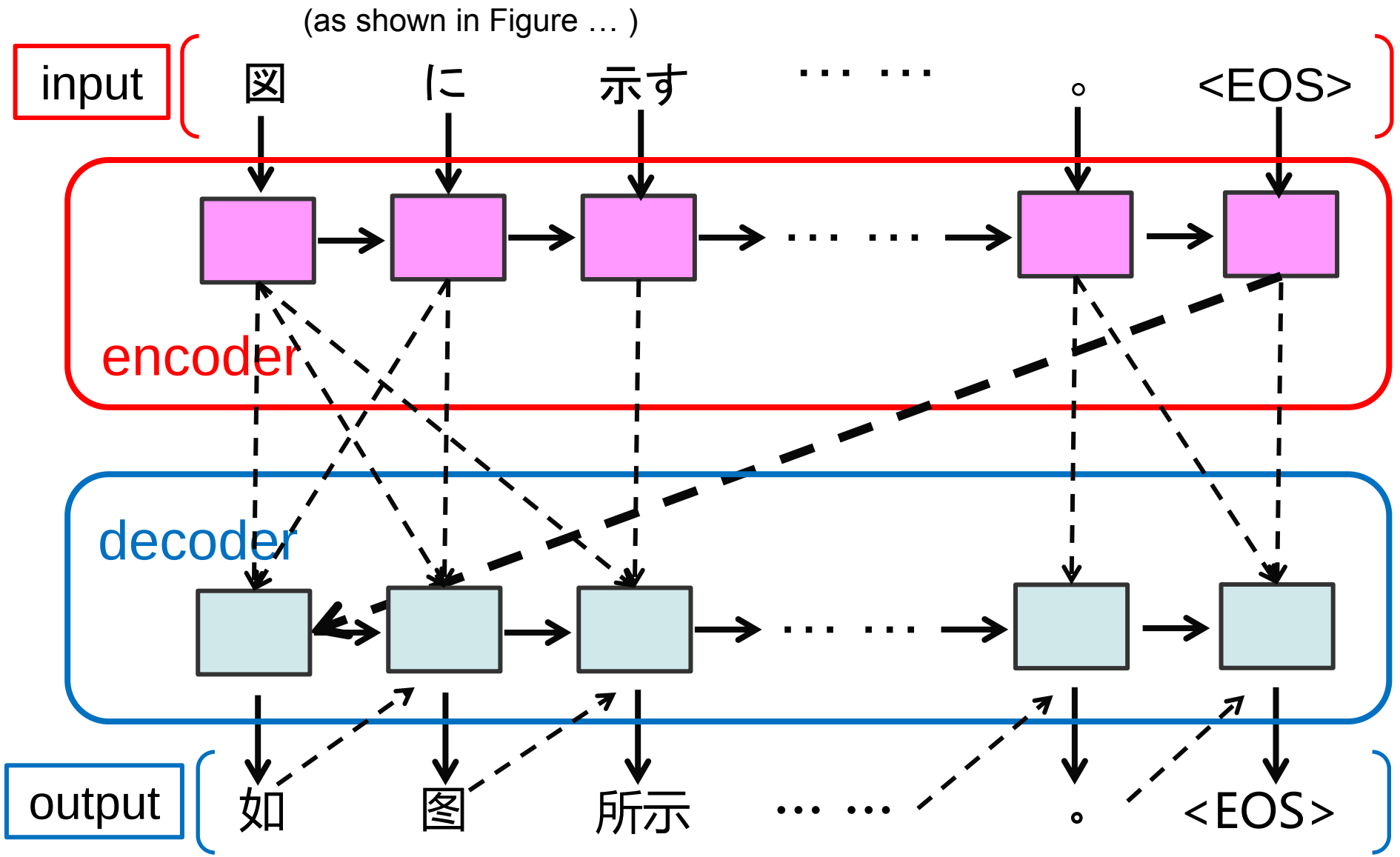
Neural Machine Translation (encoder-decoder Model and attention mechanism)

Vinyals et al. Grammar as a foreign language. In Proc. NIPS, 2015



Neural Machine Translation (encoder-decoder Model and attention mechanism)

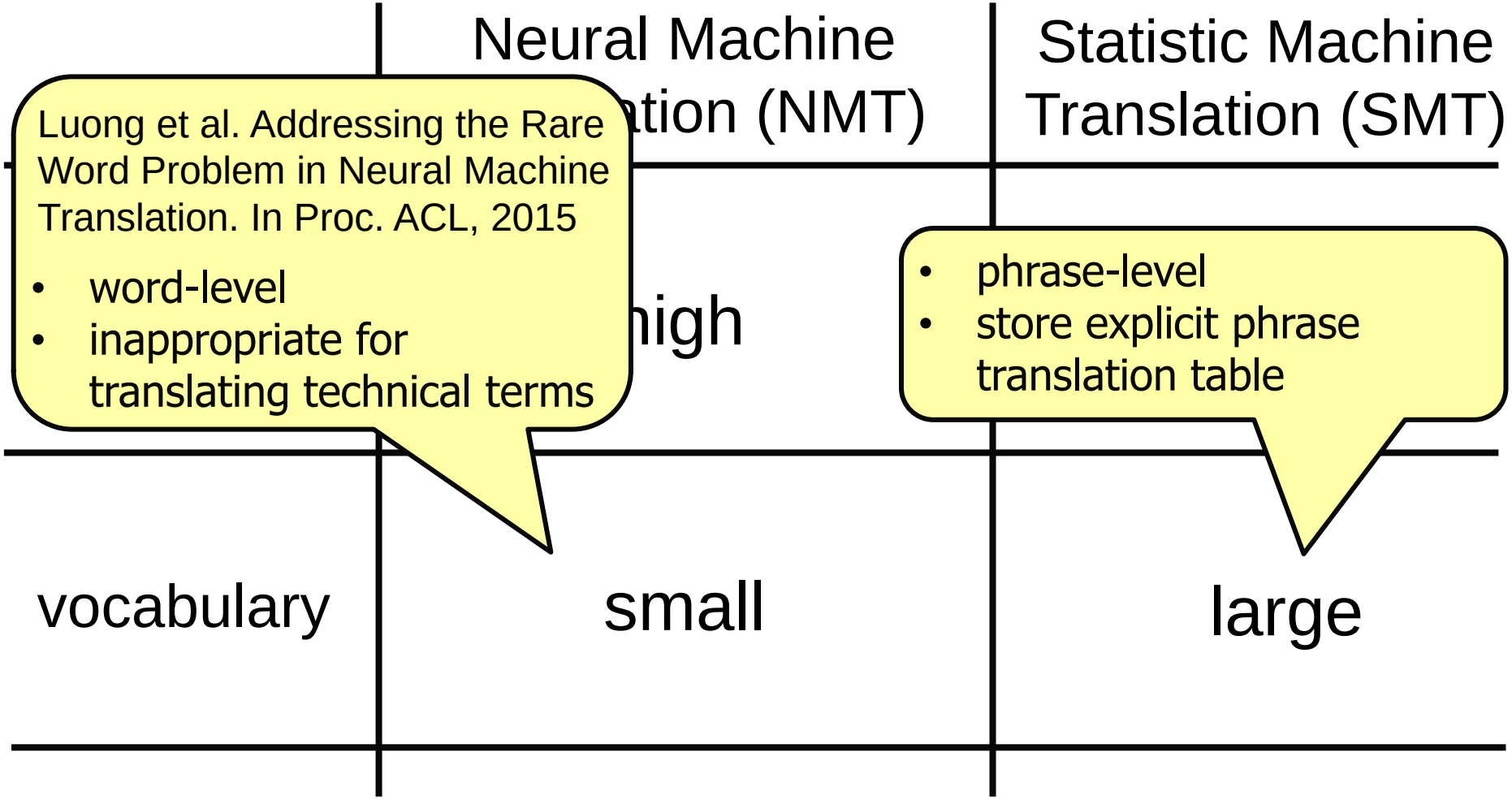
Vinyals et al. Grammar as a foreign language. In Proc. NIPS, 2015



Neural Machine Translation VS Statistic Machine Translation

	Neural Machine Translation (NMT)	Statistic Machine Translation (SMT)
fluency	high	low
vocabulary	small	large

Neural Machine Translation VS Statistic Machine Translation



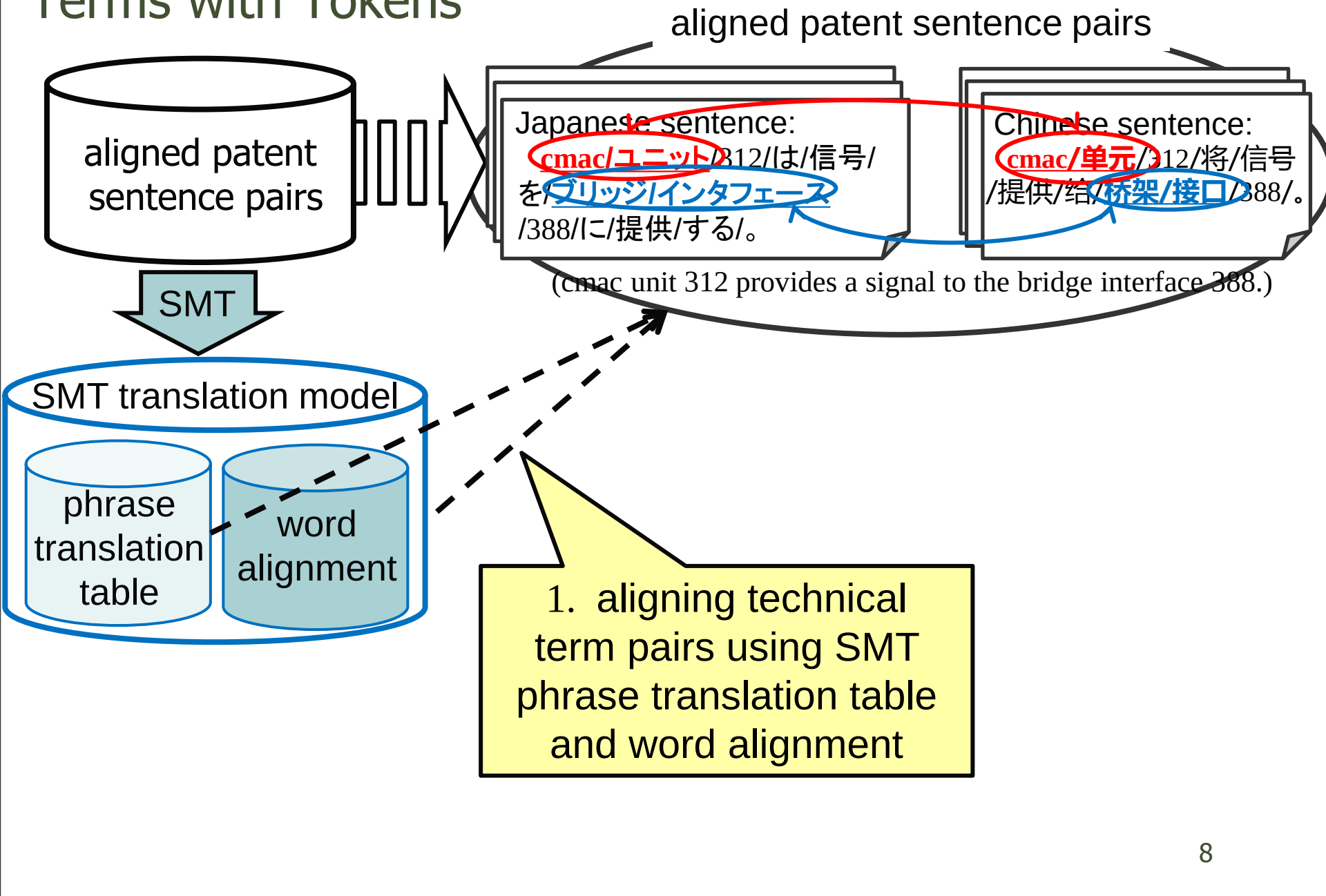
NMT with a Large Vocabulary of Technical Terms

- Step 1. training NMT model with technical term tokens
- Step 2. applying NMT model with technical term tokens
  - Approach 1. NMT decoding and SMT technical term translation
  - Approach 2. NMT rescoring of 1,000-best SMT translations (not as fluent as Approach 1)

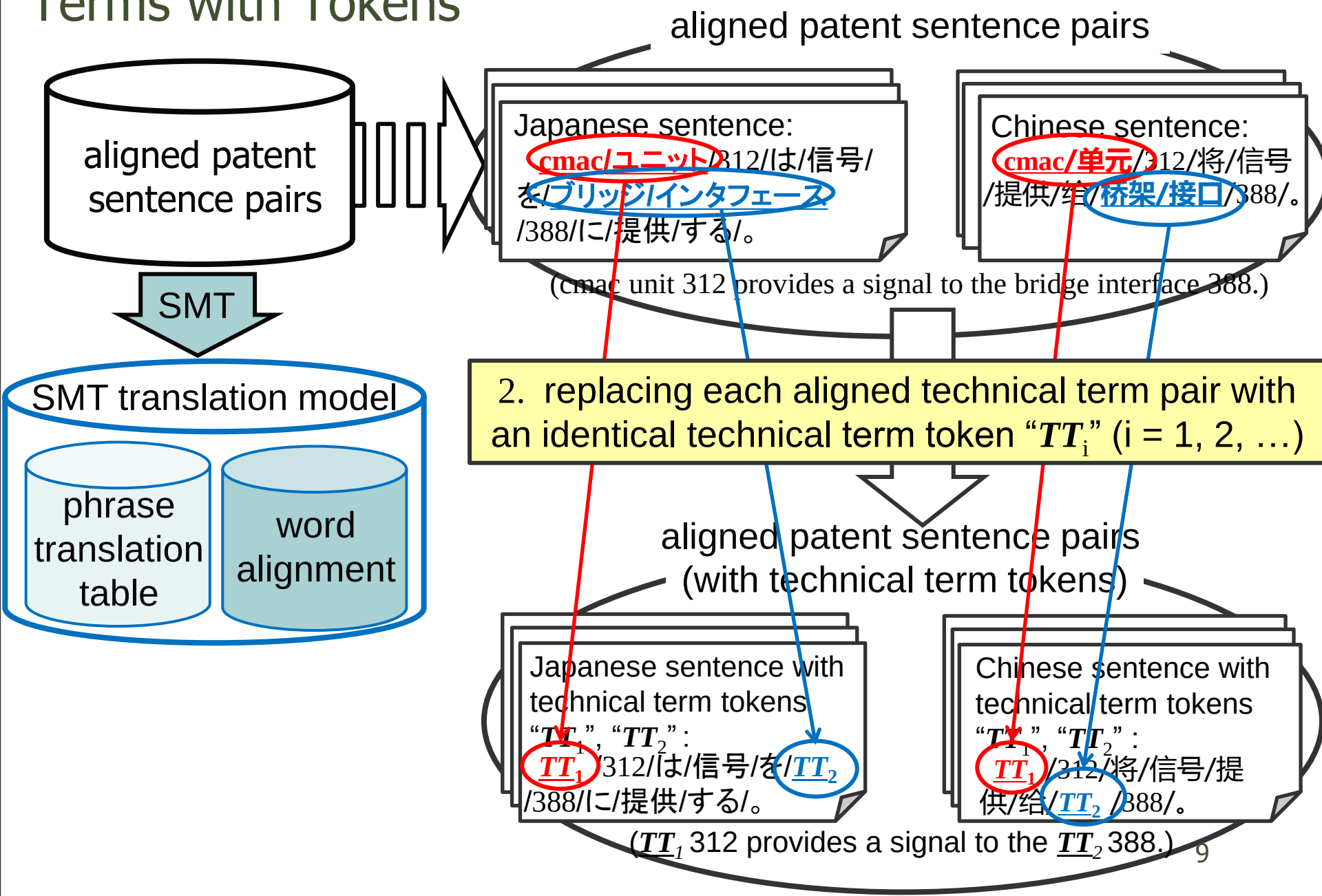
NMT with a Large Vocabulary of Technical Terms

- Step 1. training NMT model with technical term tokens
- Step 2. applying NMT model with technical term tokens
  - Approach 1. NMT decoding and SMT technical term translation
  - Approach 2. NMT rescoring of 1,000-best SMT translations (not as fluent as Approach 1)

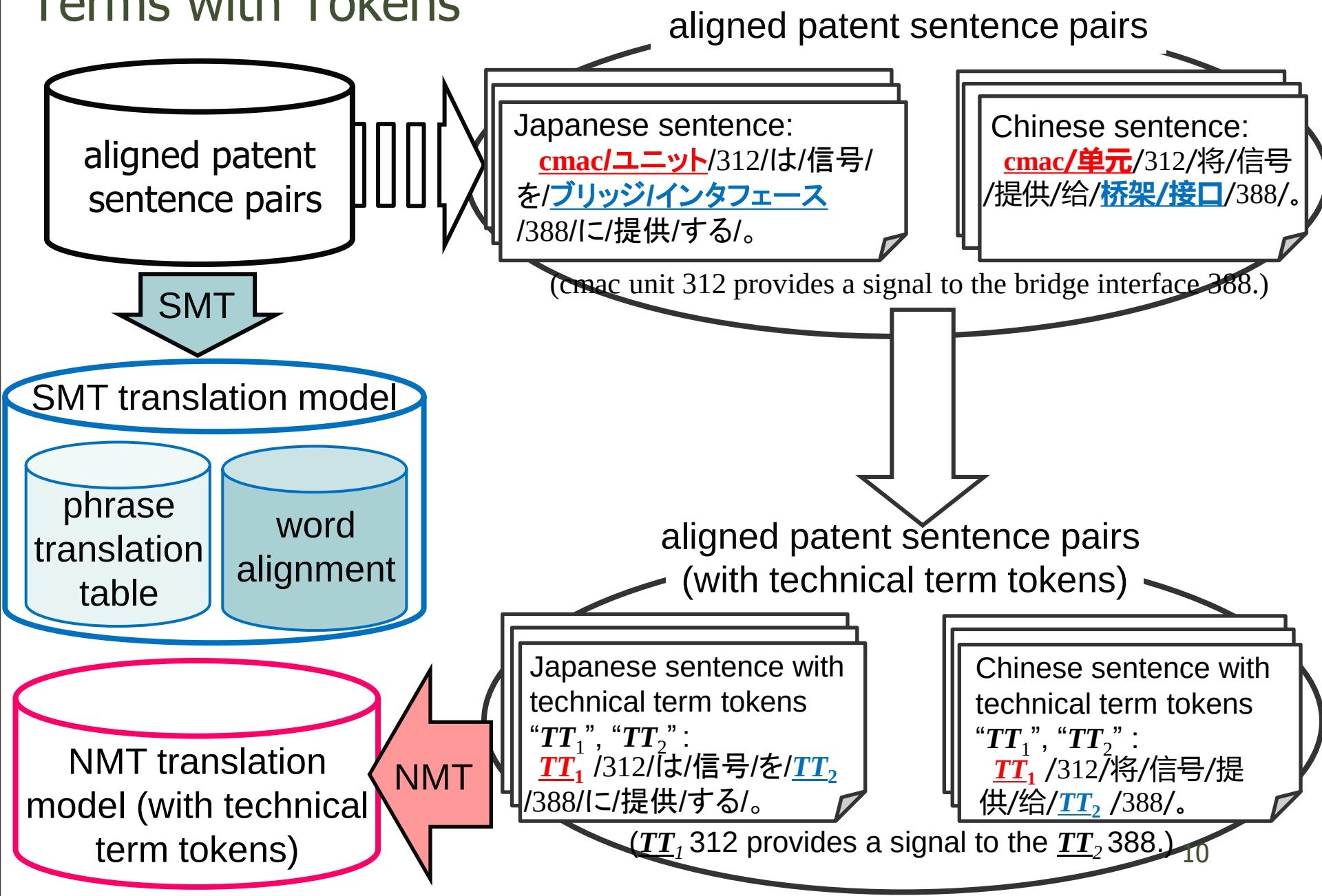
NMT training after replacing Technical Terms with Tokens



NMT training after replacing Technical Terms with Tokens



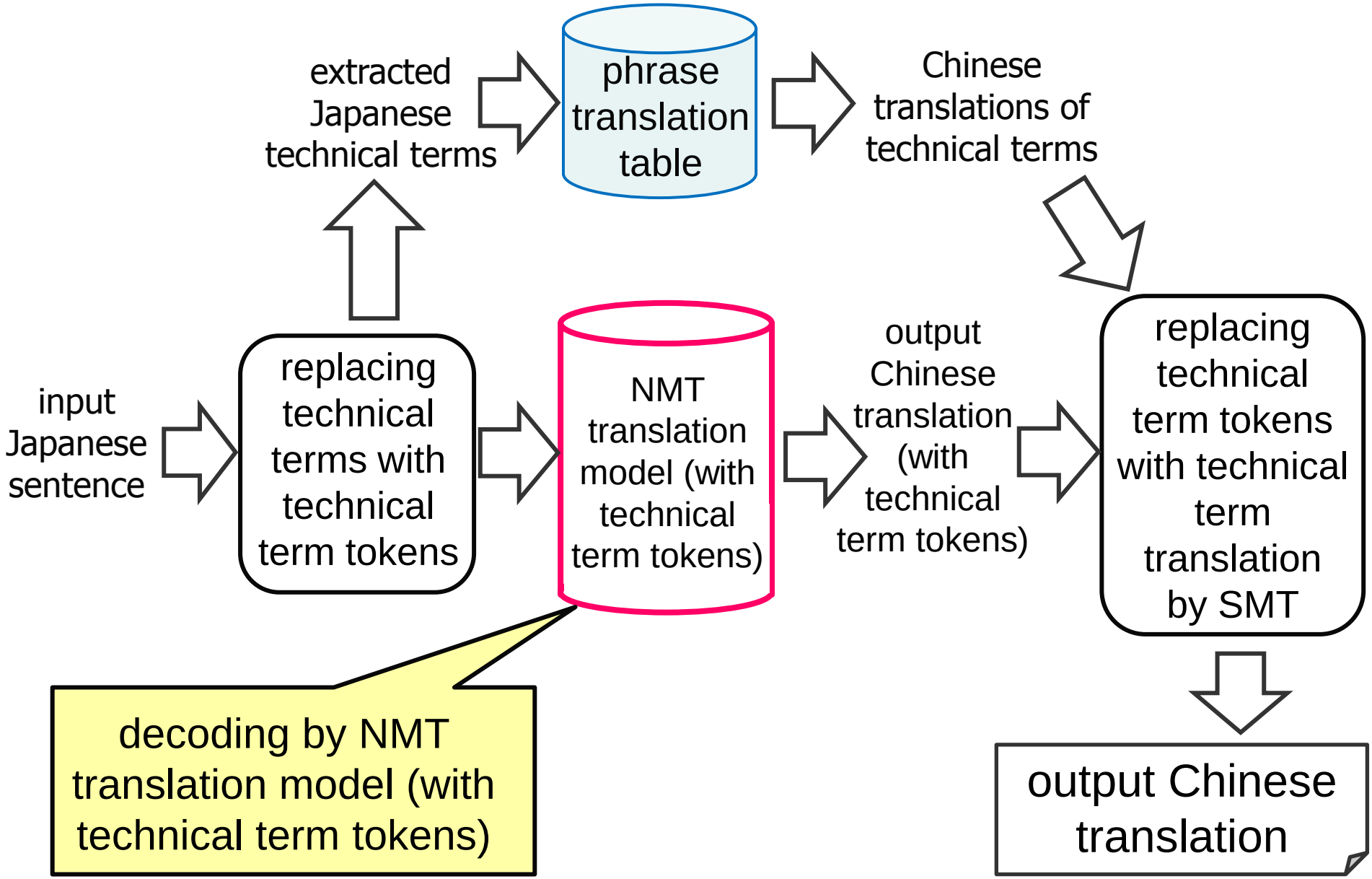
NMT training after replacing Technical Terms with Tokens



NMT with a Large Vocabulary of Technical Terms

- Step 1. training NMT model with technical term tokens
- Step 2. applying NMT model with technical term tokens
  - Approach 1. NMT decoding and SMT technical term translation
  - Approach 2. NMT rescoring of 1,000-best SMT translations (not as fluent as Approach 1)

NMT Decoding and SMT technical Term Translation



Training and Test Sets

- 2.8M parallel sentences extracted from Japanese-Chinese patent families
  - Randomly selected 1,000 sentence pairs as the test set

pair 1:
J: ･･･冷蔵庫および**冷蔵庫扉閉鎖装置**･･･
C: ･･･电冰箱及**电冰箱门锁闭装置**･･･。
(closing appliance of Refrigerator's door)

pair 2:
J: ･･･**真空断熱材**及びその製造方法･･･.
C: ･･･**真空绝热材料**及其制作方法･･･。
(vacuum thermal insulation)

⋮

pair *n*:
J: ･･･運動に関する3次元**デカルト座標系**を定める。
C: ･･･以限定用于描述运动的三维**笛卡尔坐标系**･･･。
(Cartesian coordinate system)

Experiments Settings

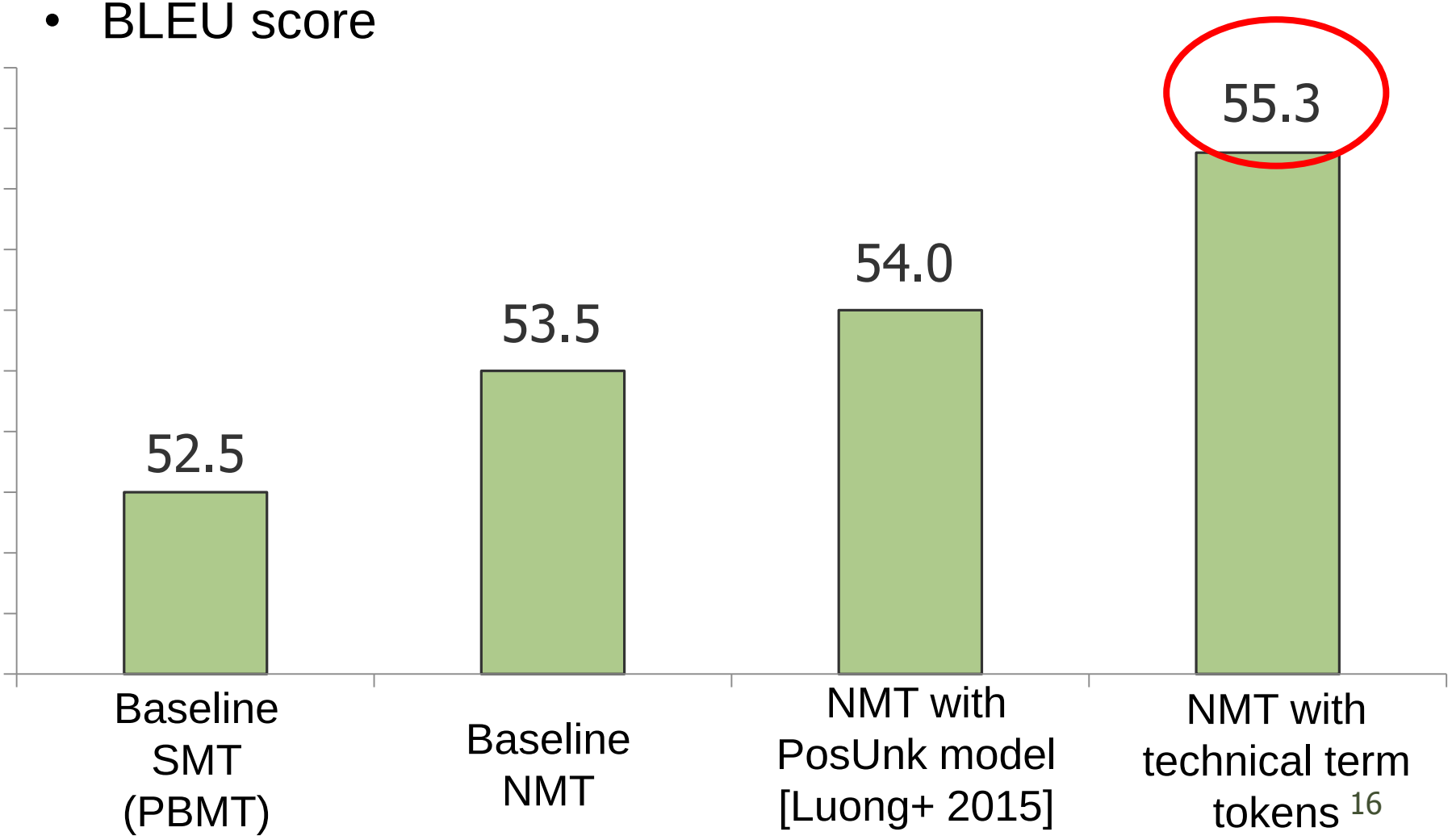
- Baseline SMT (PBMT)
  - phrase-based SMT model trained with the same training set using Moses.
- Baseline NMT
  - uni-direction model with attention mechanism
  - 3 layer deep LSTMs with 512 cells in each layer and a 512-dimensional word embedding.
  - limit both the Japanese vocabulary and the Chinese vocabulary to 40,000 most frequently used word
  - more training details.

Experiments Settings

- NMT with PosUnk model [Luong+ 2015]
  - same training parameters with the baseline NMT
  - training NMT model with PosUnk model
- NMT with technical term tokens
  - same training parameters with the baseline NMT
  - training NMT model after replacing technical terms with tokens.

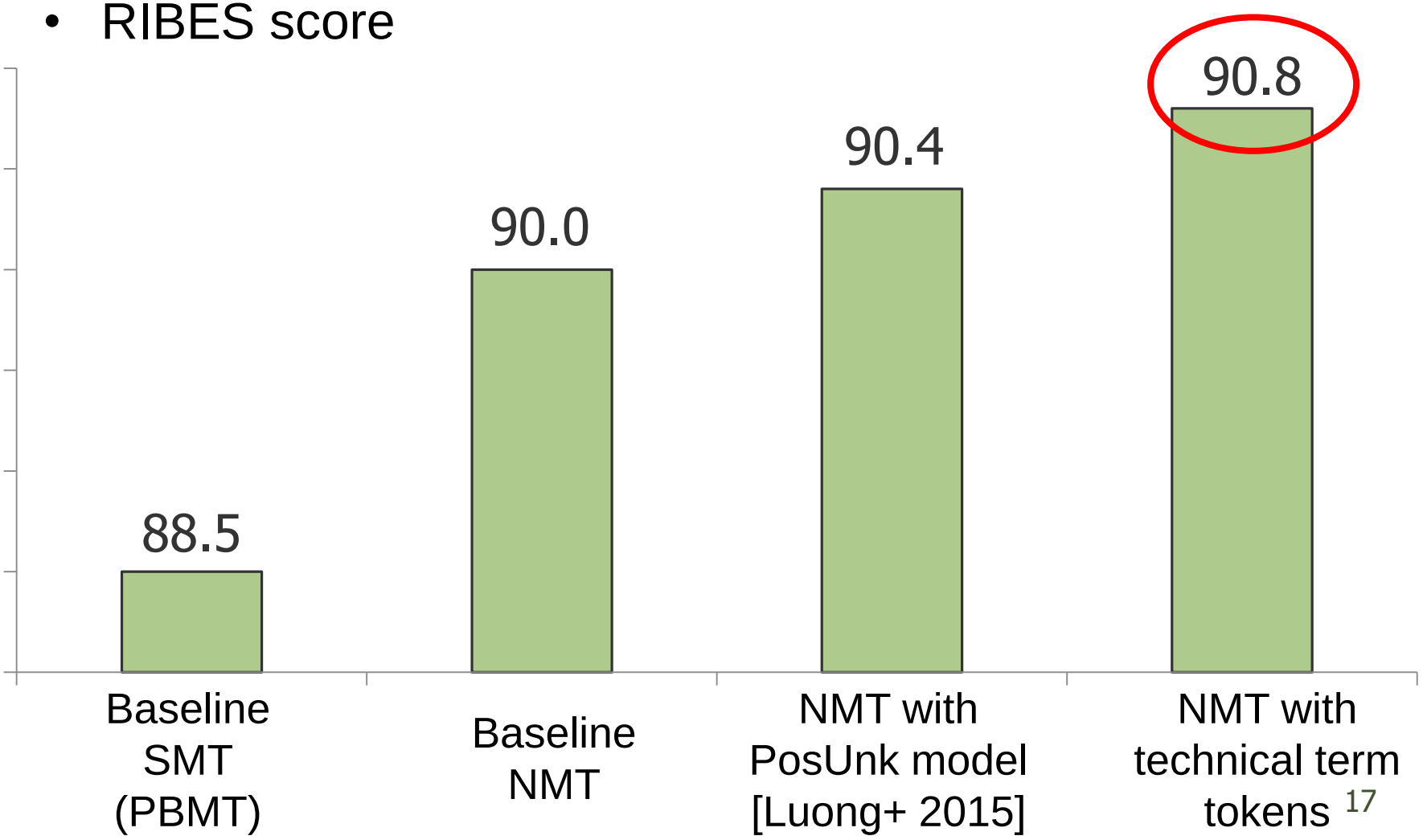
Evaluation Results (automatic evaluation - BLEU)

- BLEU score



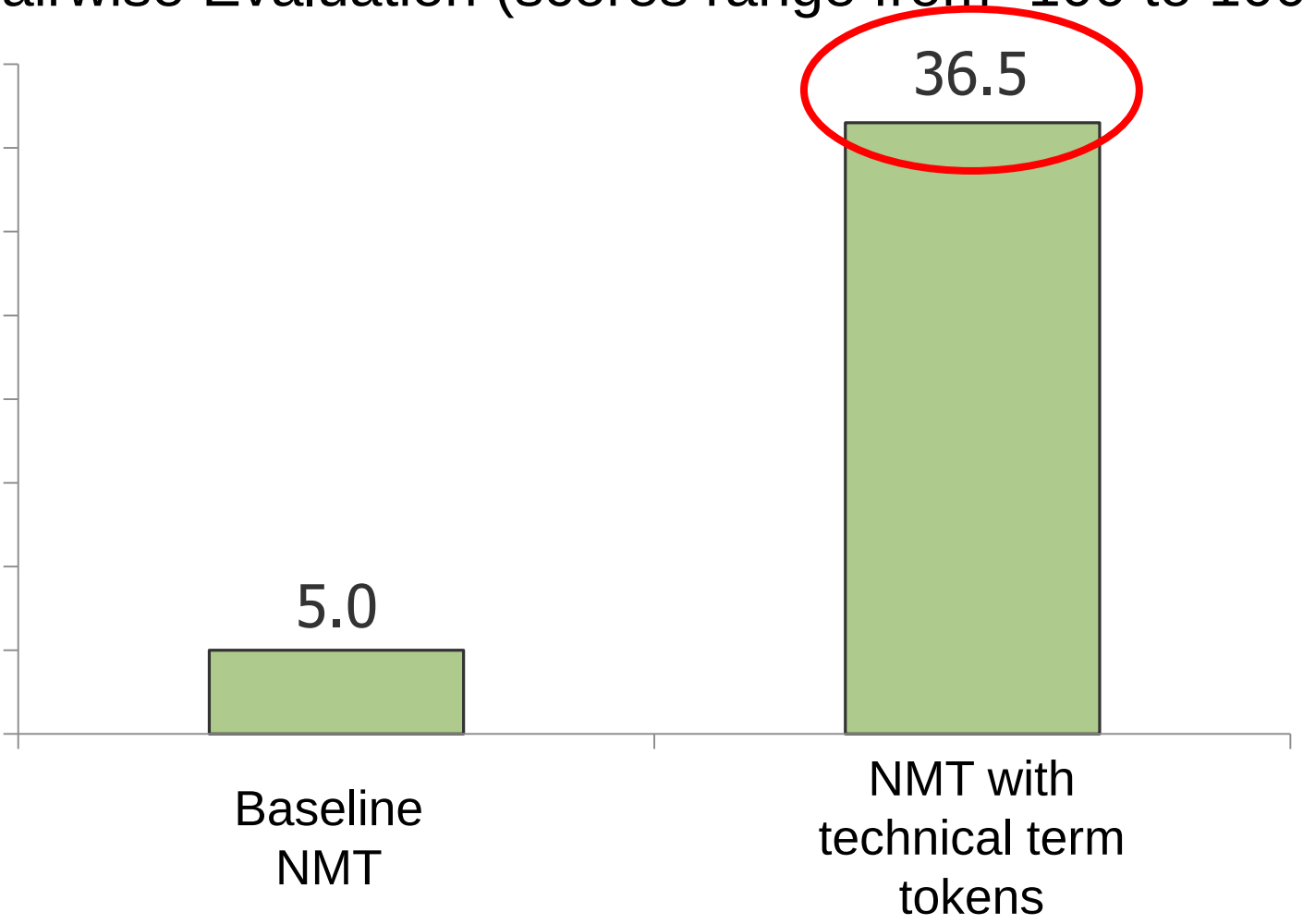
Evaluation Results (automatic evaluation - RIBES)

- RIBES score



Evaluation Results
(human evaluation – pairwise evaluation)

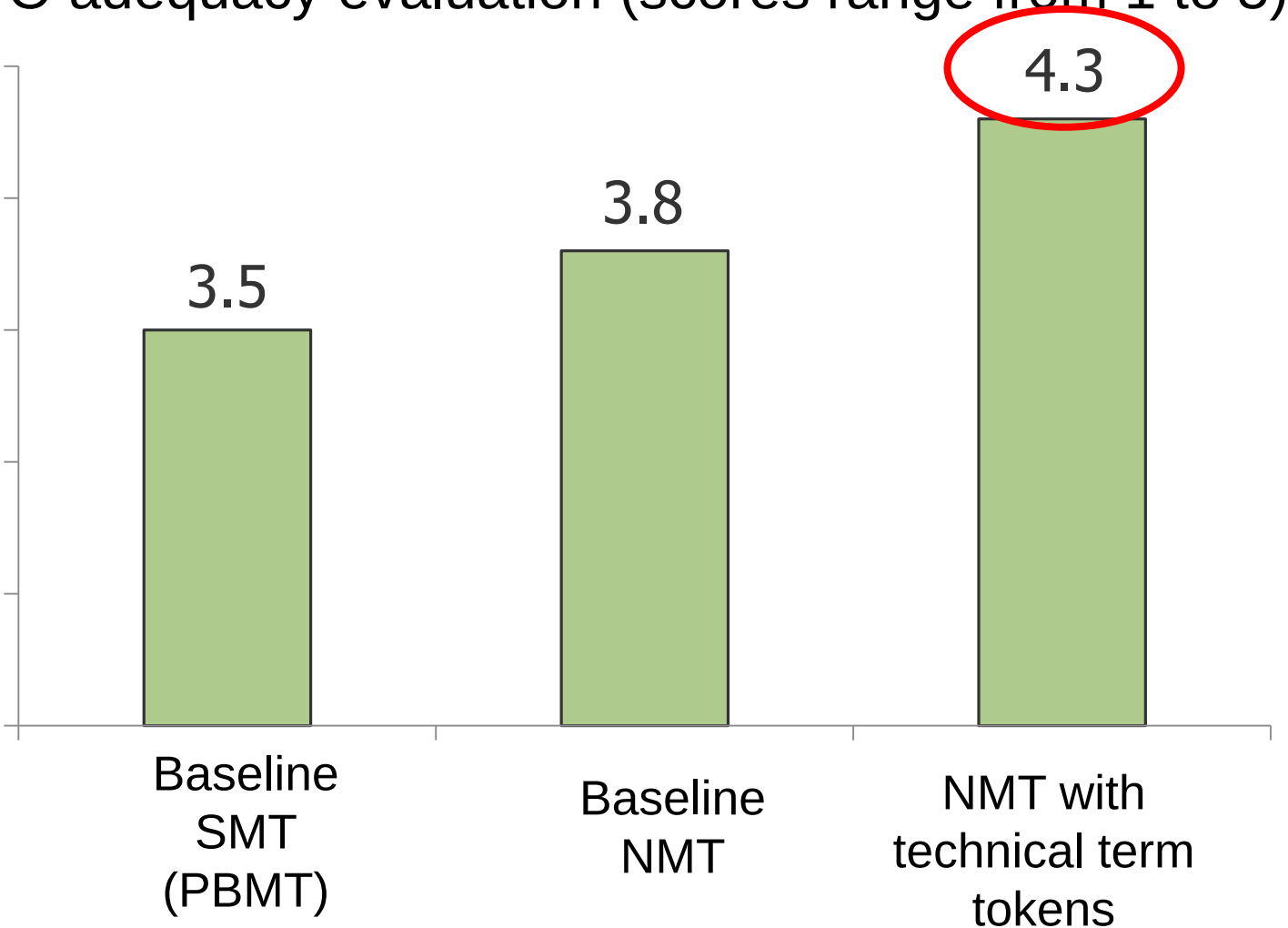
- Pairwise Evaluation (scores range from -100 to 100)



Evaluation Results

(human evaluation – JPO adequacy evaluation)

- JPO adequacy evaluation (scores range from 1 to 5)



Example of correct translations produced by NMT decoding

Japanese sentence

次に、酸化膜をhf洗浄により除去した後、**貼り合わせウェーハ**の剥離面から酸素イオンを注入した。

(Next, after removing an oxide film by hf washing, we inject oxygen ions from the peeled surface of the **laminated wafer**.)

NMT decoding
by the **PROPOSED** NMT

接着，通过hf洗涤除去氧化膜后，从**贴合晶片**的剥离面注入氧离子。

(TRANSLATION: Subsequently, after the oxide film was removed by washing with hf, and oxygen ions were injected from the peeled surface of the **laminated wafer**.)

Correct!

NMT decoding
by the **BASELINE** NMT

接着，通过hf清洗除去氧化膜后，从**贴合的UNK**的剥离面注入氧气。

(TRANSLATION: Then, after the oxide film was removed by hf cleaning, oxygen was injected from the peeled surface of the **laminated UNK**.)

Incorrect!

The Chinese word “晶片”(wafer) is out of the vocabulary

Example of correct translations produced by NMT decoding

Japanese sentence

次に、酸化膜をhf洗浄により除去した後、**貼り合わせウェーハ**の剥離面から酸素イオンを注入した。

(Next, after removing an oxide film by hf washing, we inject oxygen ions from the peeled surface of the **laminated wafer**.)

Japanese technical term “**貼り合わせウェーハ**”(laminated wafer) is translated as Chinese technical term “贴合晶片” by SMT

NMT decoding by the **PROPOSED** NMT

接着，通过hf洗涤除去氧化膜后，从**贴合晶片**的剥离面注入氧离子。

(TRANSLATION: Subsequently, after the oxide film was removed by washing with hf, and oxygen ions were injected from the peeled surface of the **laminated wafer**.)

Correct!

NMT decoding by the **BASELINE** NMT

接着，通过hf清洗除去氧化膜后，从**贴合的UNK**的剥离面注入氧气。

(TRANSLATION: Then, after the oxide film was removed by hf cleaning, oxygen was injected from the peeled surface of the **bonded UNK**.)

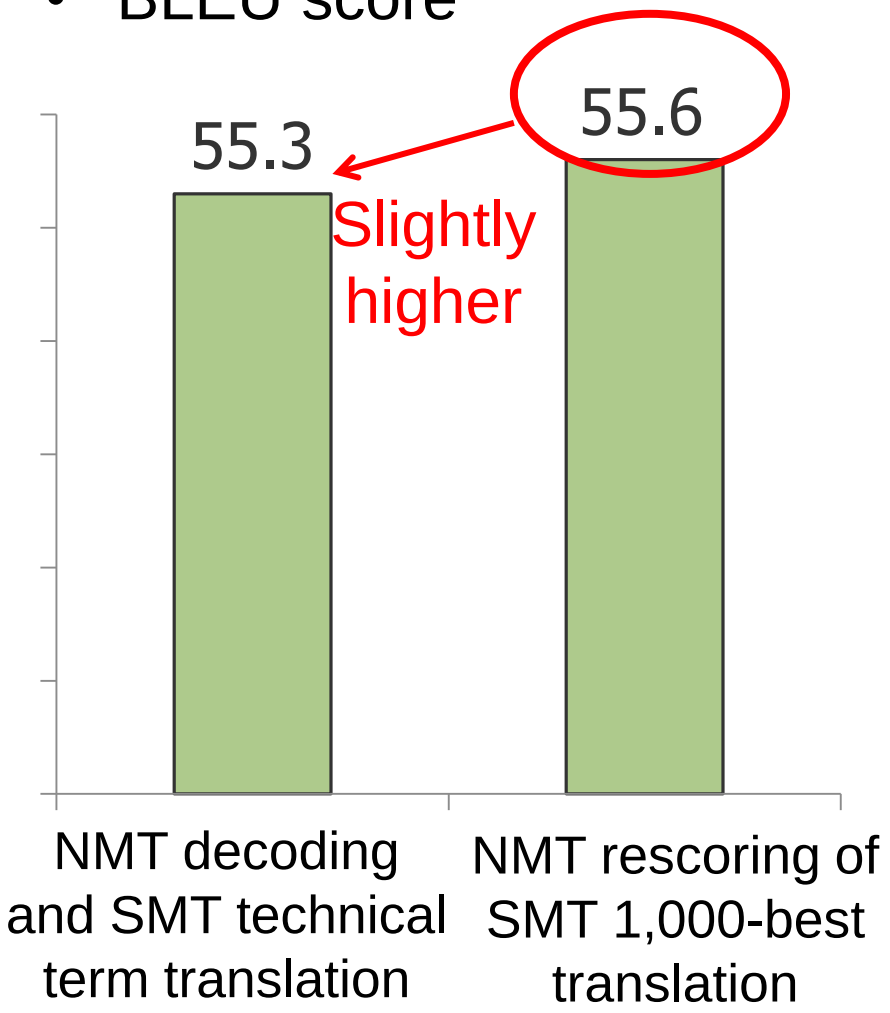
Incorrect!

NMT with a Large Vocabulary of Technical Terms

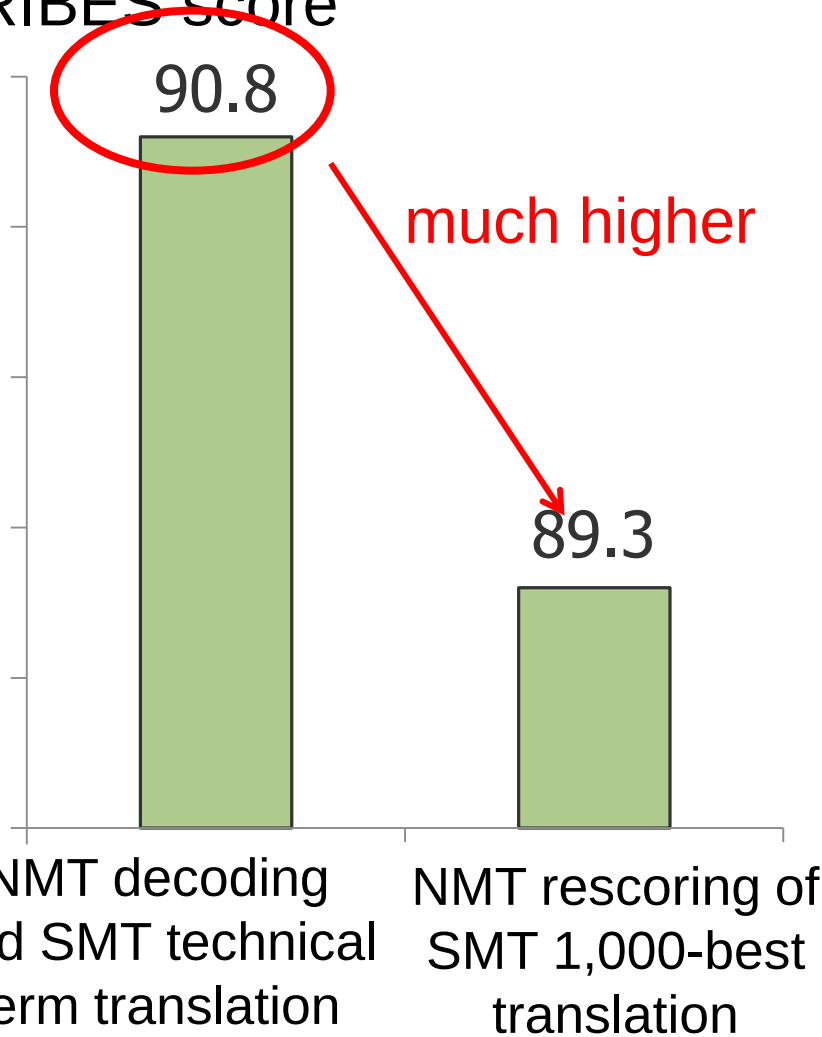
- Step 1. training NMT model with technical term tokens
- Step 2. applying NMT model with technical term tokens
  - Approach 1. NMT decoding and SMT technical term translation
  - Approach 2. NMT rescoring of 1,000-best SMT translations (not as fluent as Approach 1)

Evaluation Results (2) (automatic evaluation)

- BLEU score

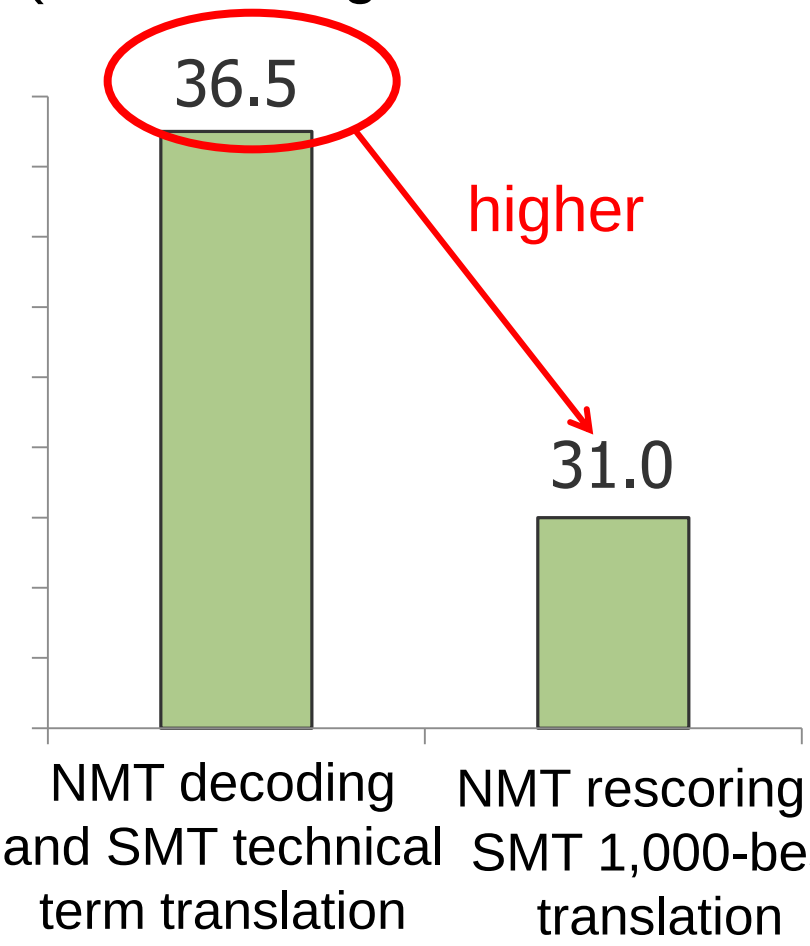


- RIBES score

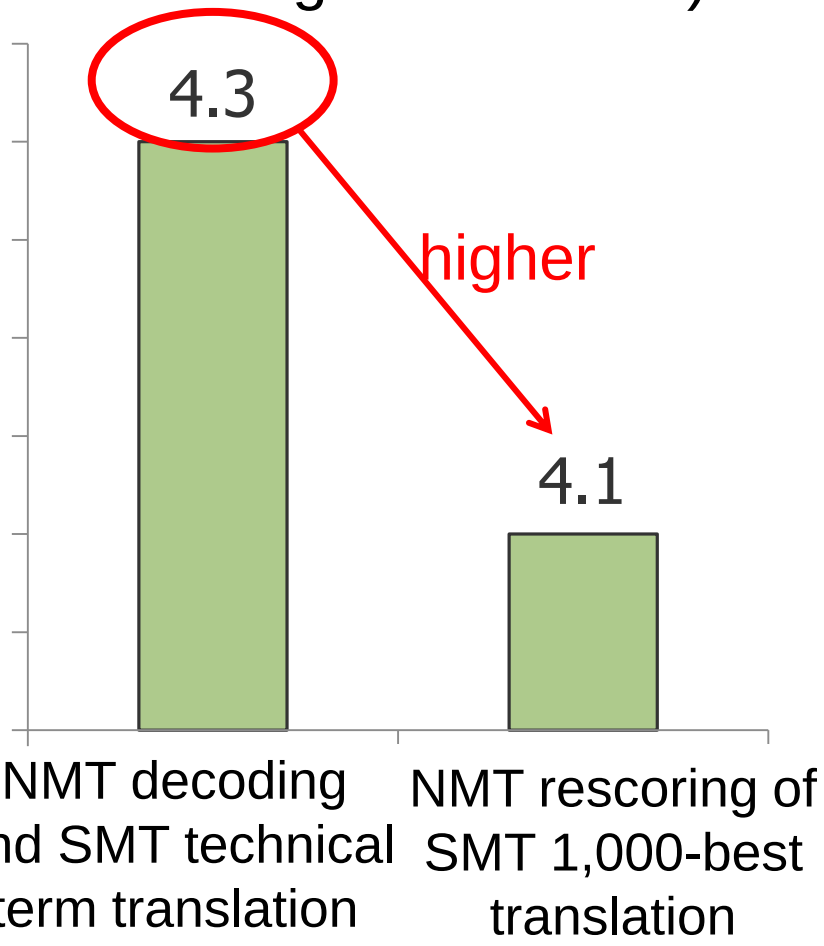


Evaluation Results (2)
(human evaluation)

- Pairwise Evaluation
(scores range from -100 to 100)



- JPO adequacy evaluation
(scores range from 1 to 5)



Conclusion and Future Work

- Translating patent sentences with a large vocabulary of technical terms by training an NMT system on a bilingual corpus, wherein technical terms are replaced with tokens
- Evaluation experiments on Japanese-Chinese patent sentences proved the effectiveness of the proposed method
- Future Work:
  - Evaluate the proposed method with a bidirection NMT system (Bahdanau et al. 2015)
  - Rescore 1,000-best NMT translations by using SMT system

Thank you for your attention!