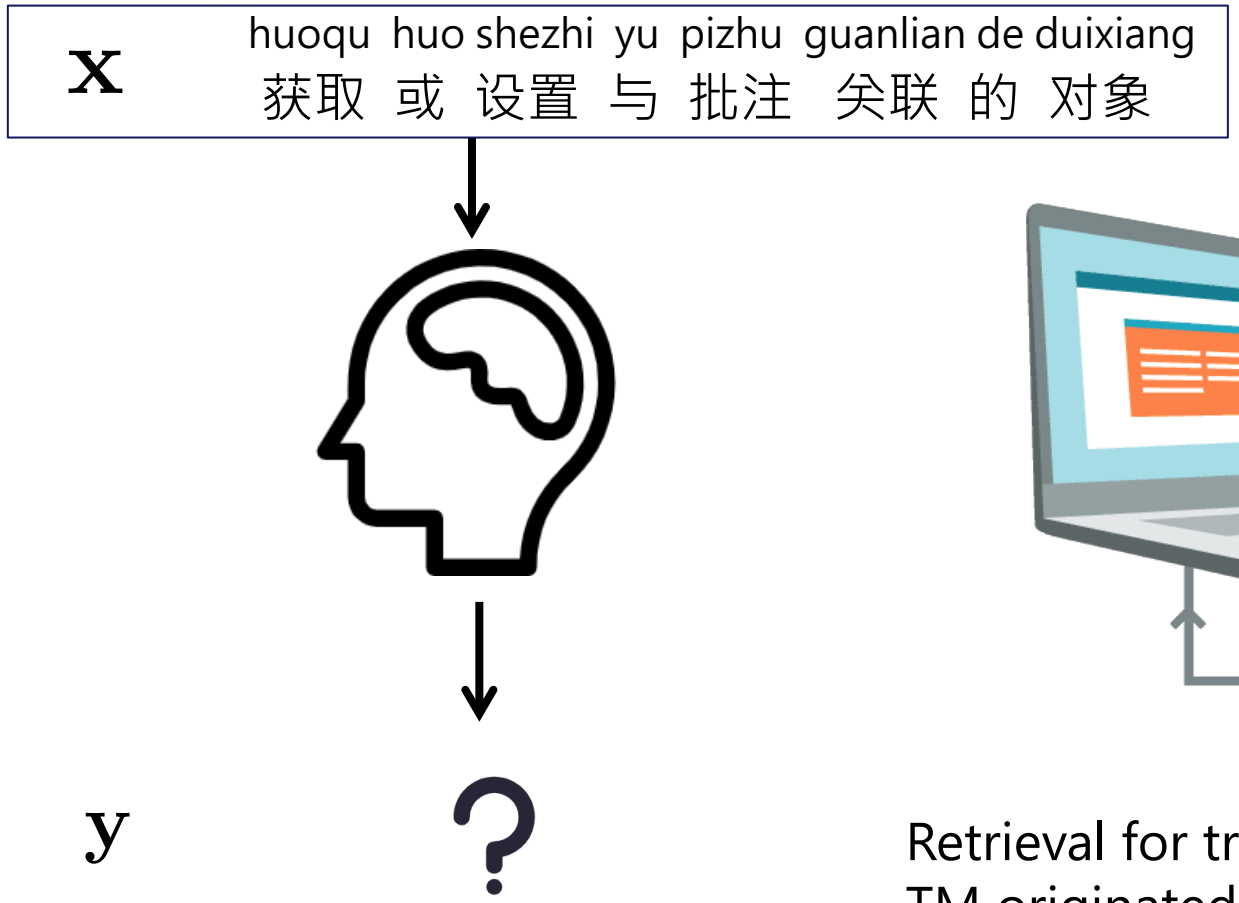
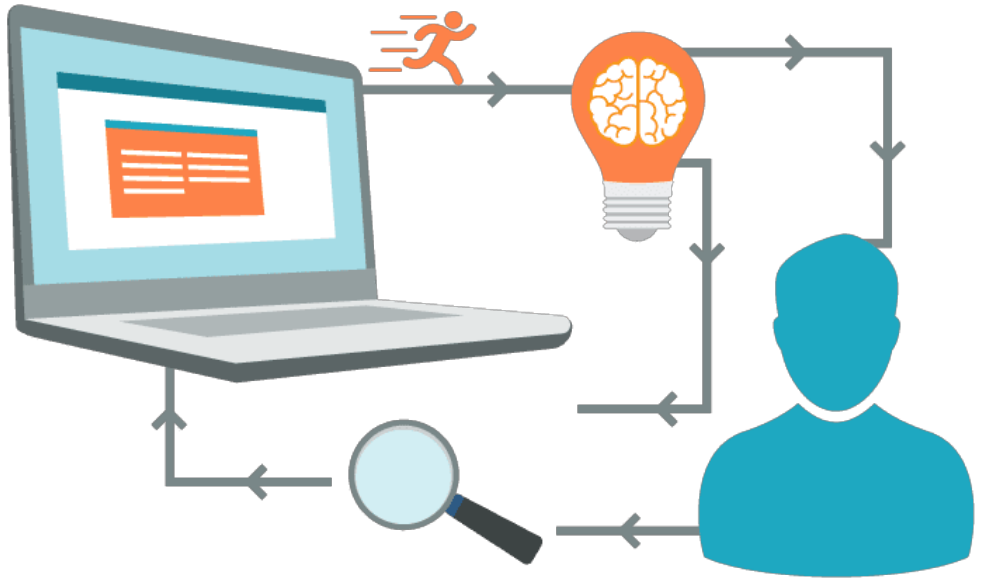
Outline

- Background and Introduction
- Language Modeling
- Open-Domain Dialogue Systems
- **Neural Machine Translation**
  - Motivation
  - TM-augmented NMT Framework
  - TM-augmented Models
    - Standard model
    - Dual model
    - Unified model
- Conclusion and Outlook

Why retrieval is beneficial to translation?

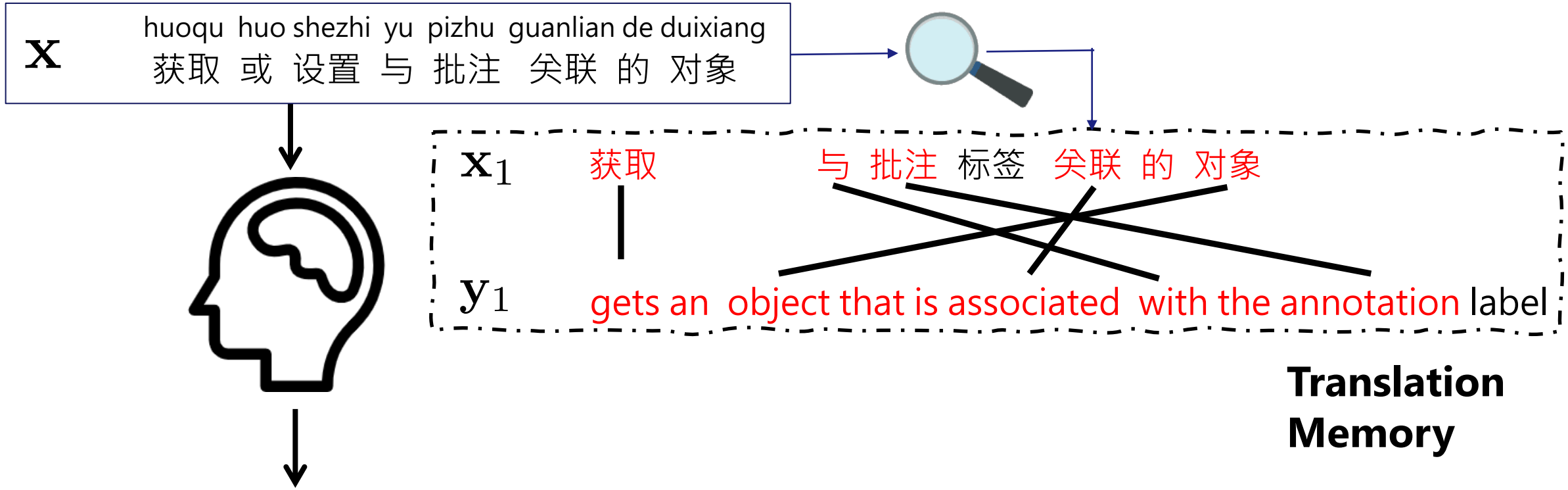




Retrieval for translation is called translation memory (TM)
TM originated from human translation scenario in 1970s

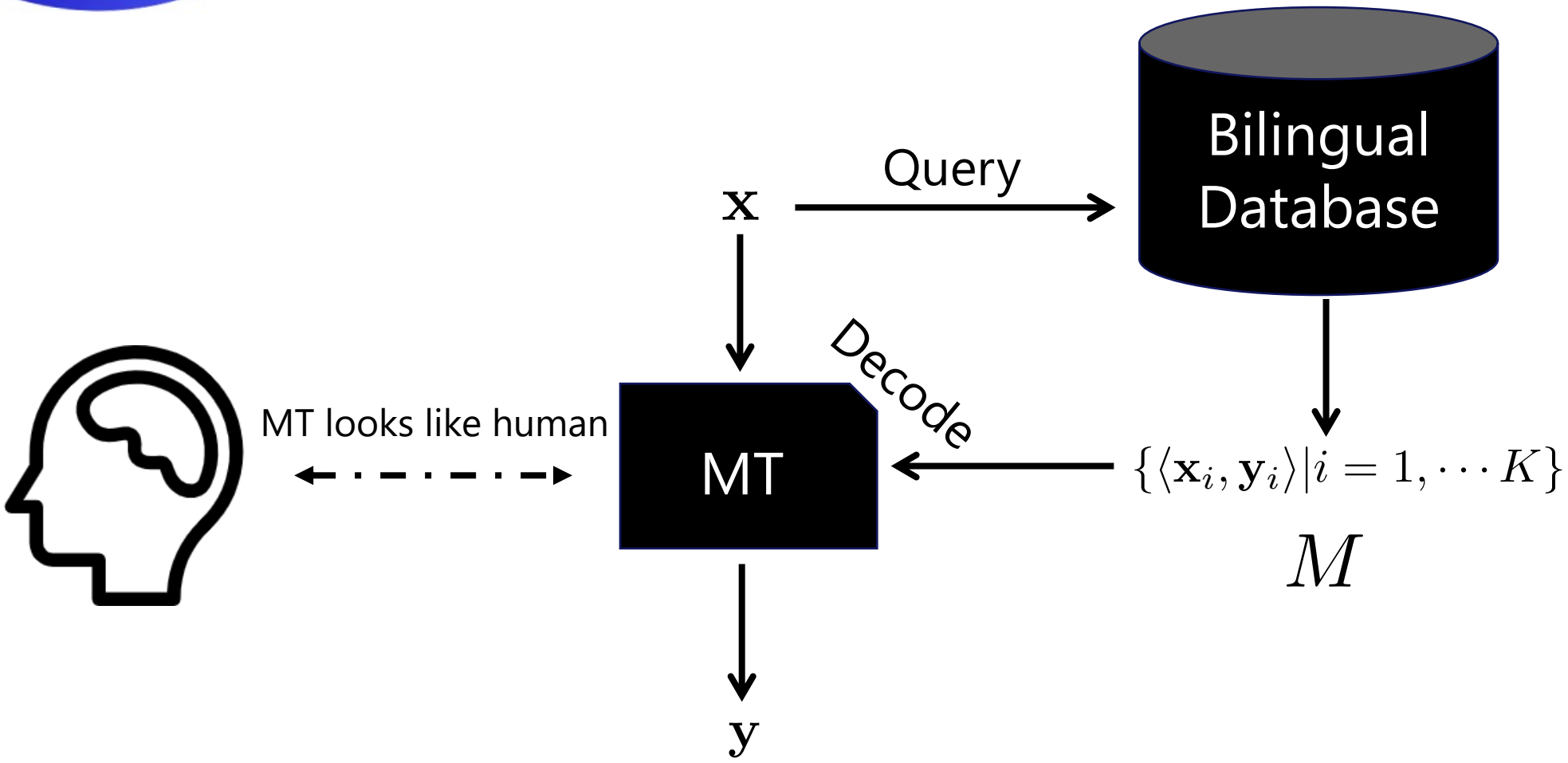
- Translating from scratch is not easy

Why retrieval is beneficial to translation?

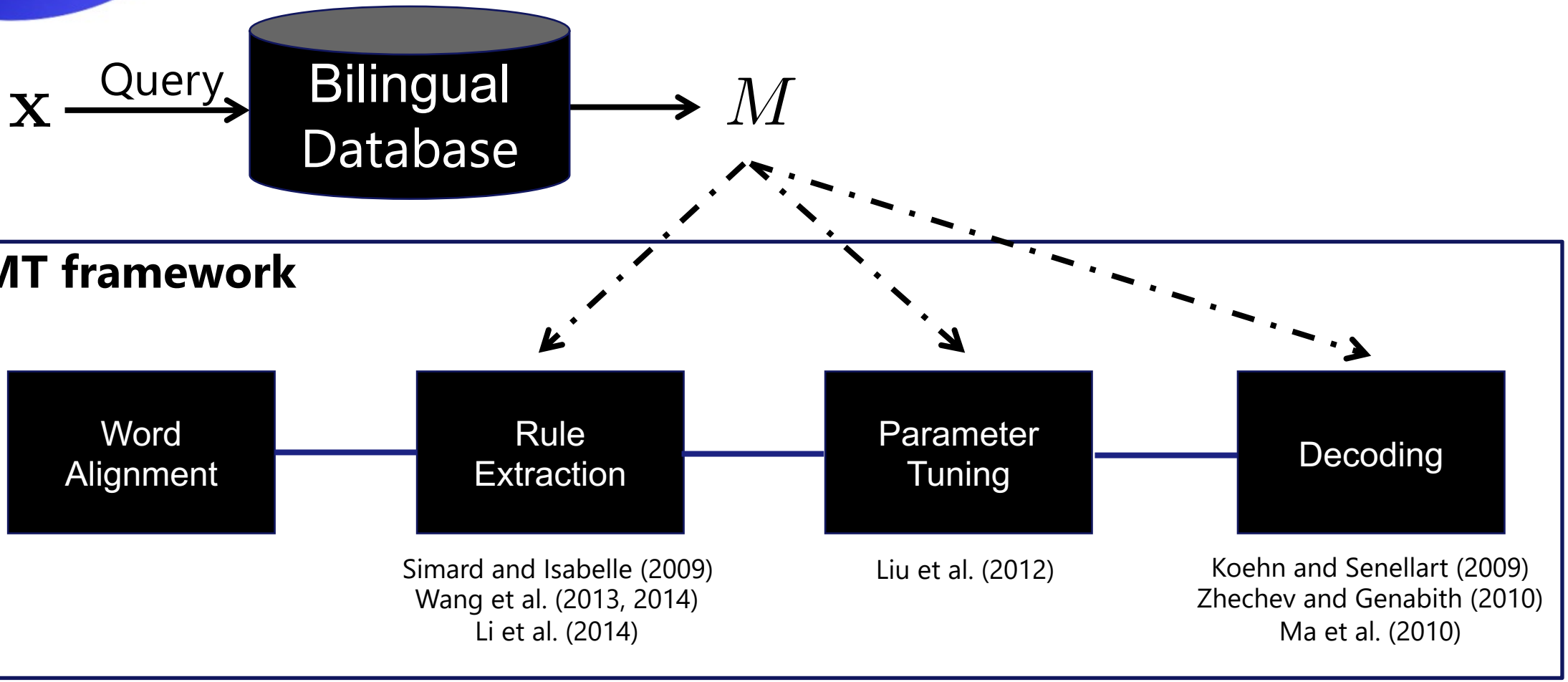


- Translation memory includes **useful translation knowledge**
- Translating from memory is easier

TM augmented MT: Paradigm



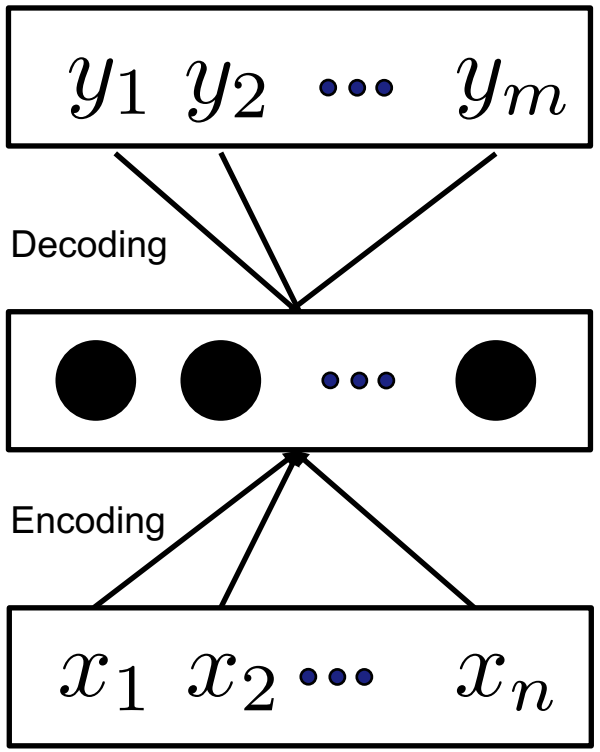
TM augmented SMT



Challenge: error propagation due to the pipeline framework

NMT: End-to-End Framework

End-to-end modeling



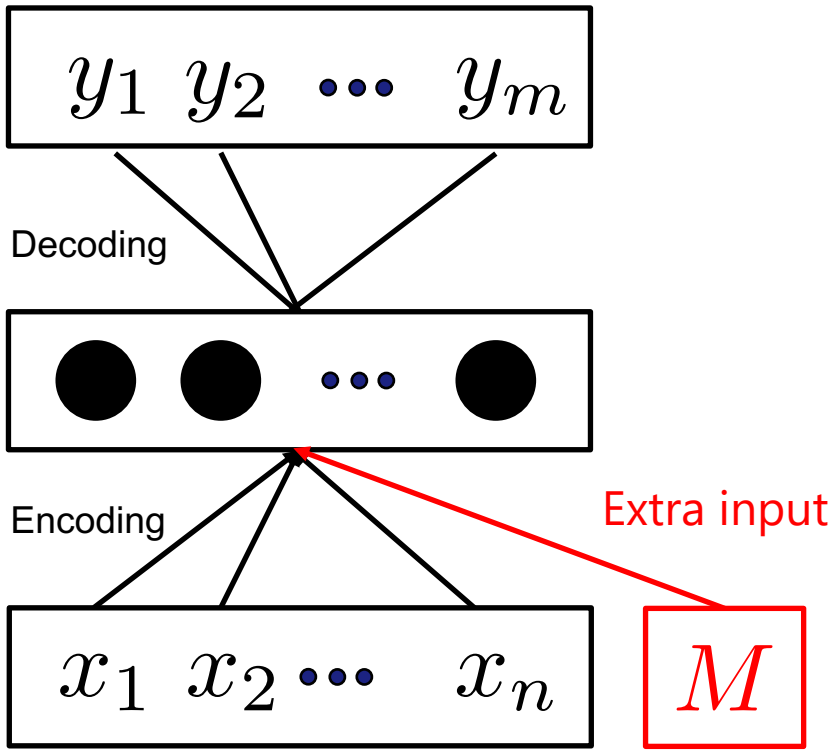
End-to-end training

$$\max_{\theta} \sum_{\langle \mathbf{x}, \mathbf{y} \rangle} \log p(\mathbf{y}|\mathbf{x}; \theta)$$

NMT achieves SOTA performance on many benchmarks

NMT: End-to-End Framework

End-to-end modeling



End-to-end training

$$\max_{\theta} \sum_{\langle \mathbf{x},\mathbf{y} \rangle} \log p(\mathbf{y}|\mathbf{x};\theta)$$

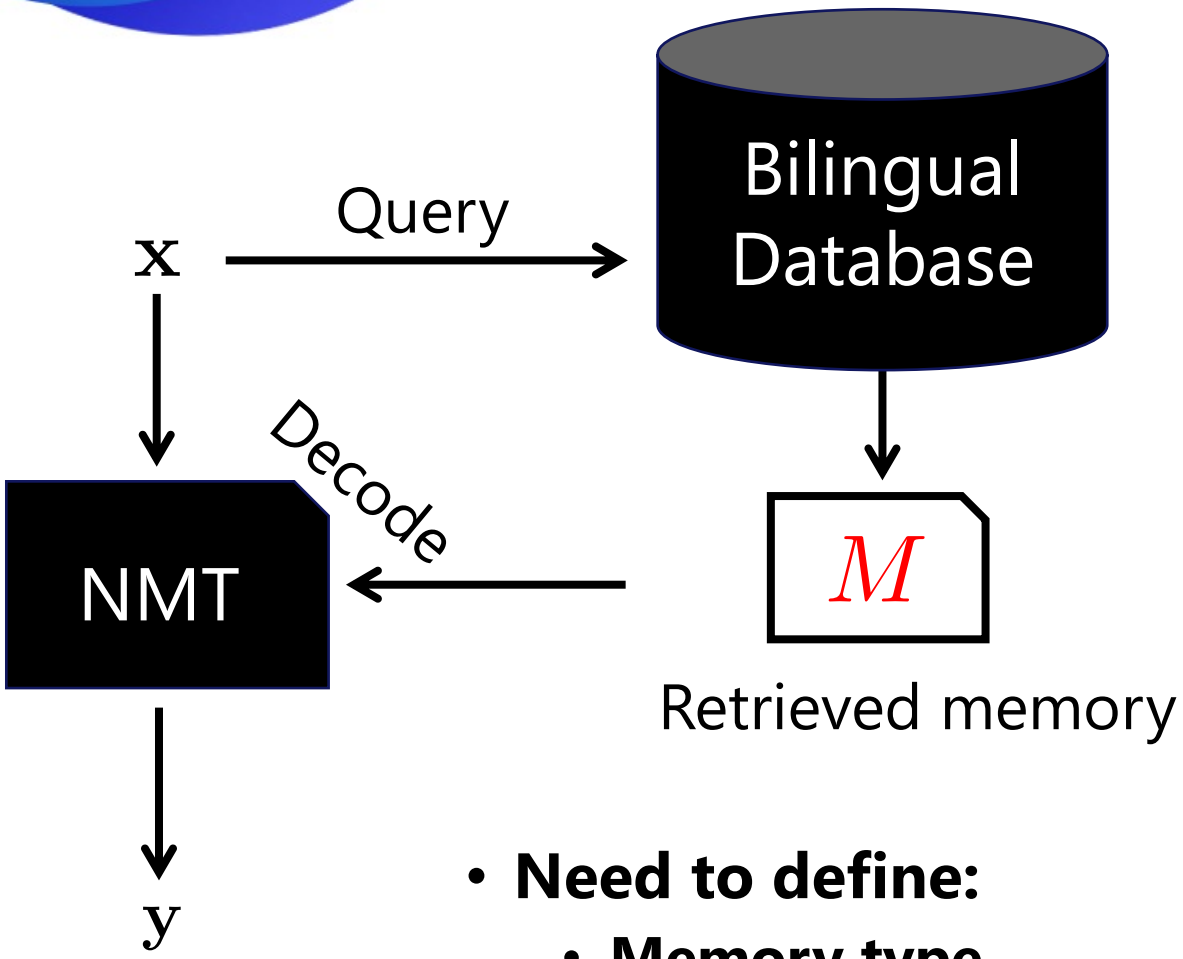
$$\max_{\theta} \sum_{\langle \mathbf{x},\mathbf{y},M \rangle} \log p(\mathbf{y}|\mathbf{x}, M;\theta)$$

Easily scaling to leverage any extra information
Making TM-augmented NMT promising

Outline

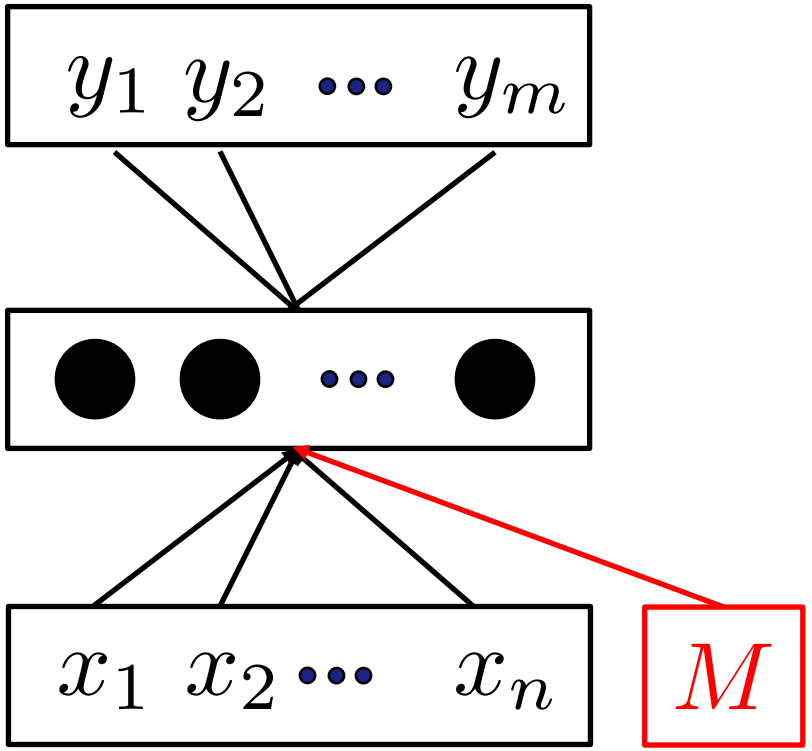
- Background and Introduction
- Language Modeling
- Open-Domain Dialogue Systems
- **Neural Machine Translation**
  - Motivation
  - **TM-augmented NMT Framework**
  - TM-augmented Models
    - Standard model
    - Dual model
    - Unified model
- Conclusion and Outlook

TM-augmented NMT Framework: Overview



- **Need to define:**
  - **Memory type**
  - **Retrieval metric**
  - **Model architecture**

End-to-end modeling



End-to-end training

$$\max_{\theta} \sum_{\langle \mathbf{x},\mathbf{y},M \rangle} \log p(\mathbf{y}|\mathbf{x}, M; \theta)$$

TM-augmented NMT Framework: Memory Type

huoqu huo shezhi yu pizhu guanlian de duixiang

$\mathbf{x}$ 获取 或 设置 与 批注 关联 的 对象

$\hat{\mathbf{y}}_{1:7}$ gets or sets an object that is ?

Test sentence

- Type 1: <sentence, word>

Query $\mathbf{x}$

$\langle \mathbf{x}^1, \mathbf{y}^1 \rangle$

Key-value pairs

$\mathbf{x}^1$ huoqu yu pizhu biaoqian guanlian de duixiang

获取 与 批注 标签 关联 的 对象

$\mathbf{y}^1$ gets an object that is **associated** with the annotation label

A sentence in database

- Type 2: <sentence, word>

Query $\mathbf{x}||\hat{\mathbf{y}}_{1:7}$

$\langle \mathbf{x}^1||\mathbf{y}^1_{1:5}, \text{associated} \rangle$

. . .

Key-value pairs

TM-augmented NMT Framework: Retrieval Metrics

huoqu huo shezhi yu pizhu guanlian de duixiang

$\mathbf{x}$ 获取 或 设置 与 批注 关联 的 对象

$\hat{\mathbf{y}}_{1:7}$ gets or sets an object that is ?

Test sentence

$\mathbf{x}^1$ huoqu yu pizhu biaoqian guanlian de duixiang

获取 与 批注 标签 关联 的 对象

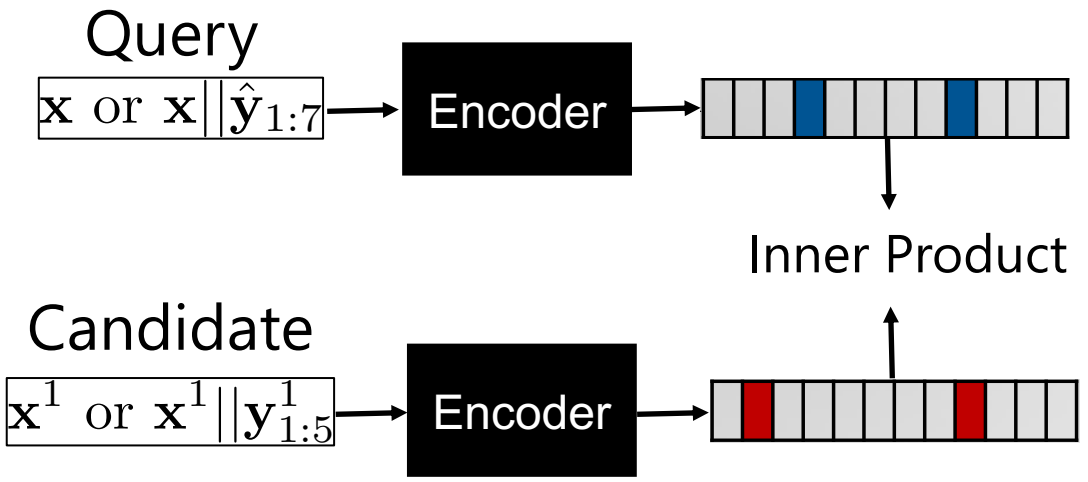
$\mathbf{y}^1$ gets an object that is **associated** with the annotation label

A sentence in database

- Word Matching
  - TF-IDF
  - Normalized edit distance

$$1 - \frac{\text{edit-dist}(\mathbf{x}, \mathbf{x}^1)}{\max(|\mathbf{x}|, |\mathbf{x}^1|)}$$

- Dense Retrieval



TM-augmented NMT: Categories

Tencent AI Lab

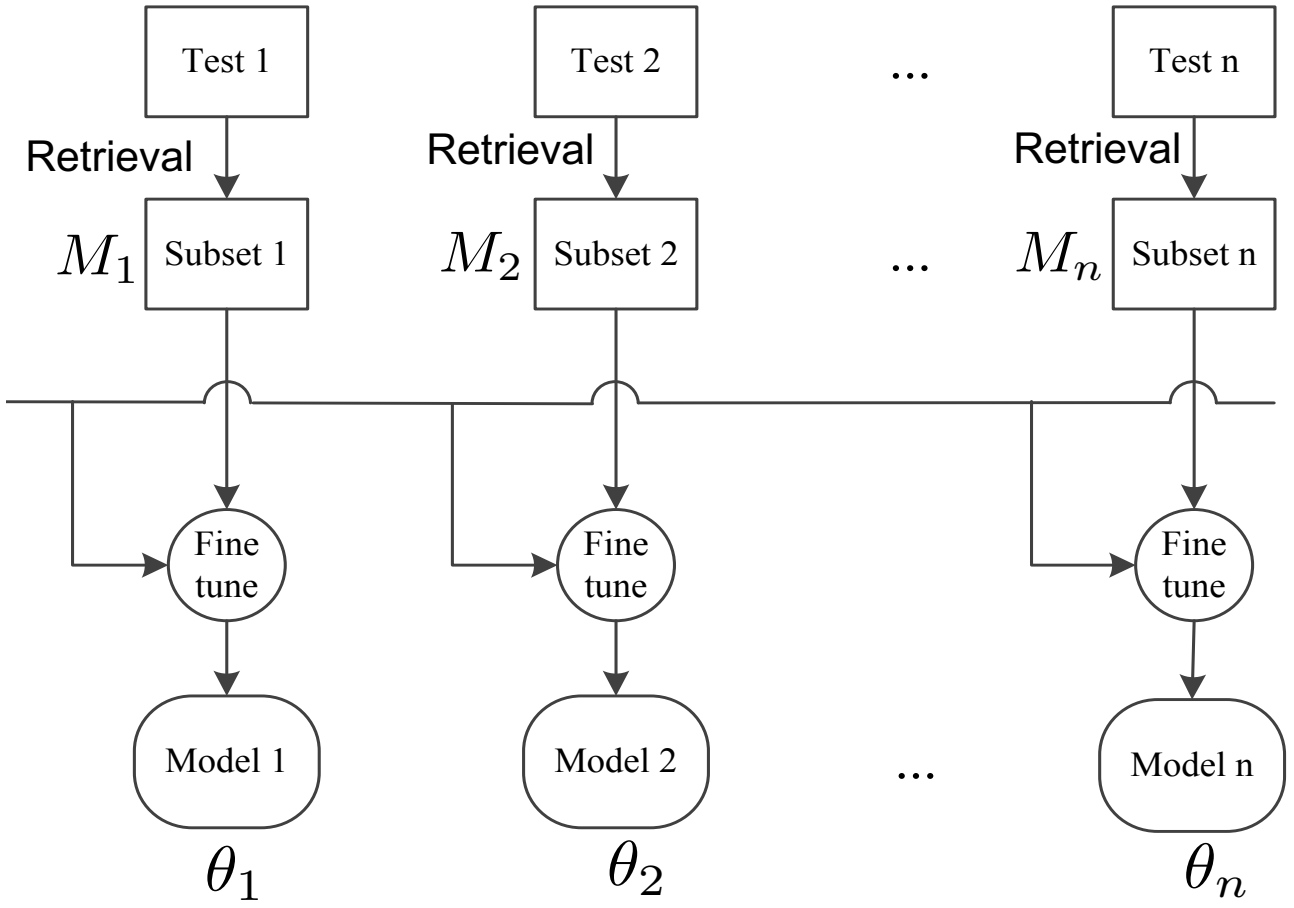
Ref.	Memory Type	Retrieval Metric	*Model Architecture*
Li et al. (2016) Farajian et al. (2017) **Bulte et al. (2019)**	<sentence, sentence>	Word Matching	***Standard model (fixed NMT architecture)***
Xu et al. (2020)	<sentence, sentence>	Word Matching Dense retrieval	
Zhang et al. (2018)	<sentence, sentence>	Word Matching	***Dual model (partially changed architecture)***
Khandelwal et al. (2021) Zheng et al. (2021) Wang et al. (2022) Meng et al. (2022)	<sentence, word>	Dense retrieval	
Gu et al. (2018) *Xia et al. (2019)* *He et al. (2021)*	<sentence, sentence>	Word Matching	***Unified model (changed architecture)***
Cai et al. (2021)	<sentence, sentence>	Dense retrieval	

Outline

Tencent
AI Lab

- Background and Introduction
- Language Modeling
- Open-Domain Dialogue Systems
- **Neural Machine Translation**
  - Motivation
  - TM-augmented NMT Framework
  - **TM-augmented Models**
    - **Standard model**
    - Dual model
    - Unified model
- Conclusion and Outlook

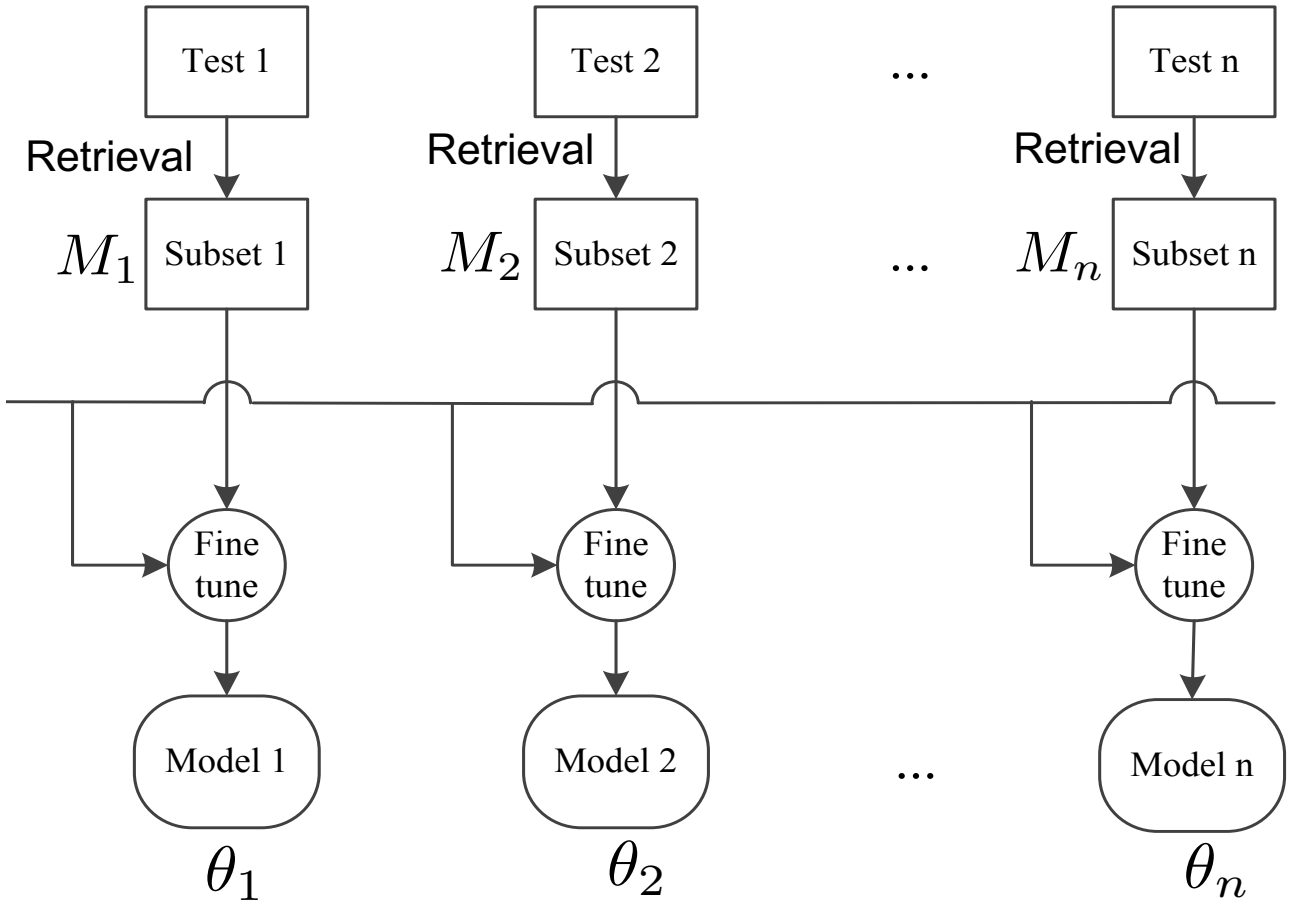
Standard Model: Finetuning



Standard NMT model
(RNN, Transformer)

Fig credit: Xiaoqing Li, Jiajun Zhang, Chengqing Zong. One sentence one model for neural machine translation. arxiv16.

Standard Model: Finetuning



Finetuning Objective

$$\max_{\theta_n} \sum_{\langle x,y\rangle \in M_n} \log p(y|x;\theta_n)$$

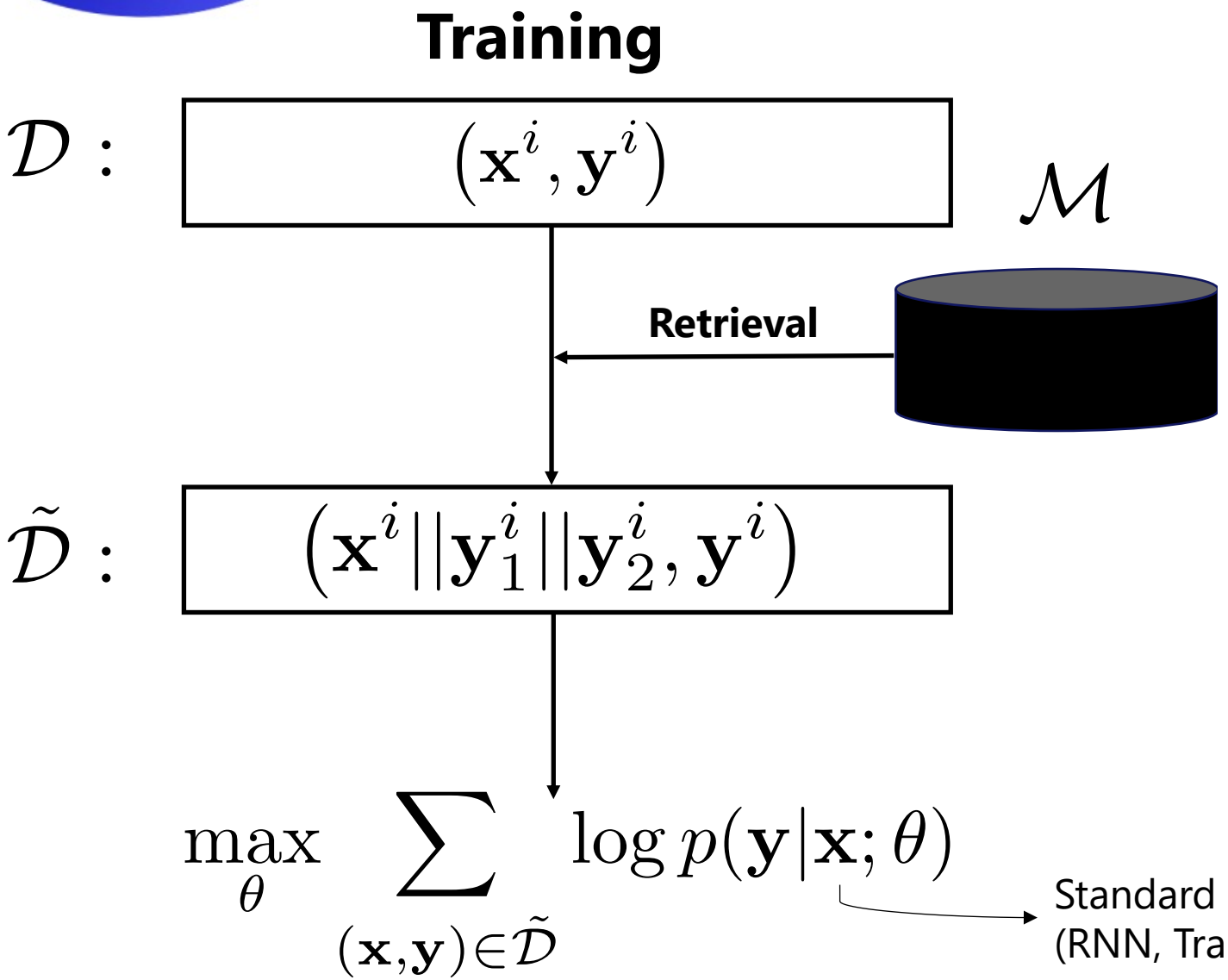
Standard NMT model (RNN, Transformer)

- Optimize θ_n
  - Run SGD on M_n
- Decode with θ_n

On-the-fly finetuning and testing

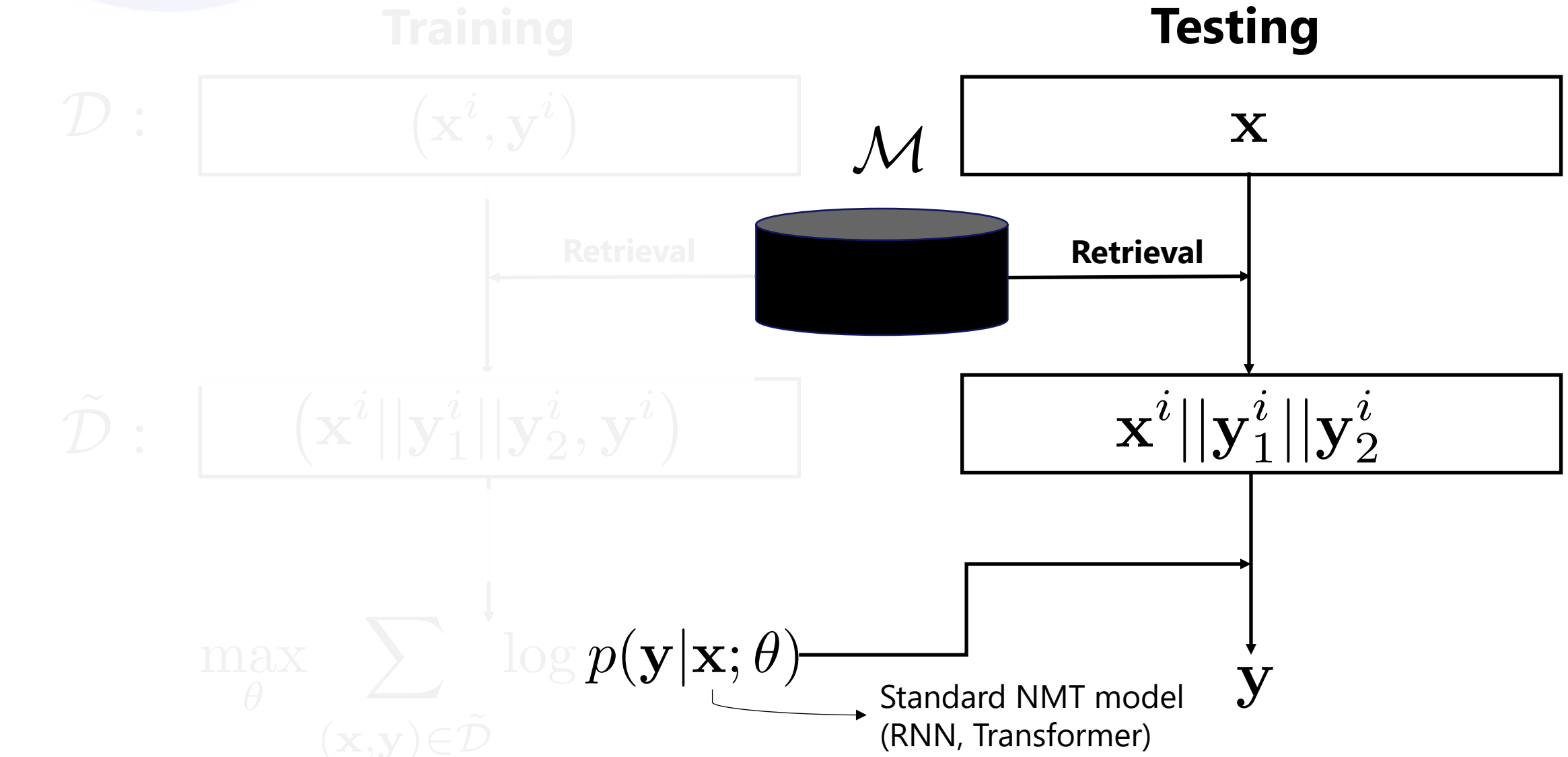
Fig credit: Xiaoqing Li, Jiajun Zhang, Chengqing Zong. One sentence one model for neural machine translation. arxiv16.

Standard Model: Input Augmentation



Bram Bulte, Arda Tezcan. Neural Fuzzy Repair: Integrating Fuzzy Matches into Neural Machine Translation. ACL19.

Standard Model: Input Augmentation



Bram Bulte, Arda Tezcan. Neural Fuzzy Repair: Integrating Fuzzy Matches into Neural Machine Translation. ACL19.

Outline

- Background and Introduction
- Language Modeling
- Open-Domain Dialogue Systems
- **Neural Machine Translation**
  - Motivation
  - TM-augmented NMT Framework
  - **TM-augmented Models**
    - Standard model
    - **Dual model**
    - Unified model
- Conclusion and Outlook

Dual model

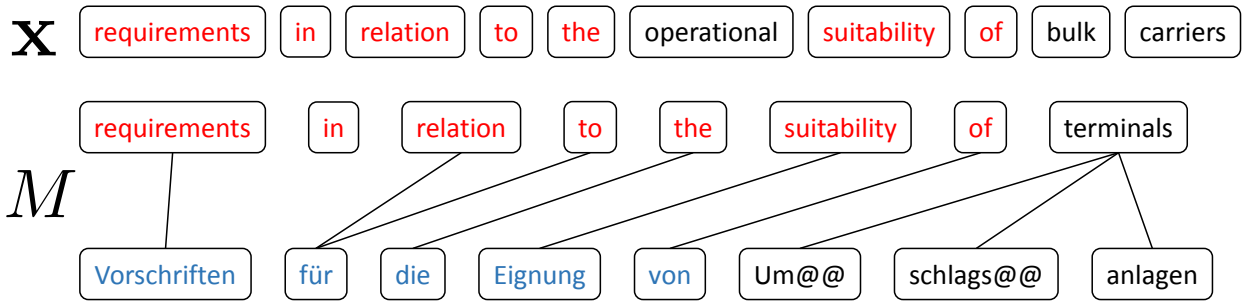
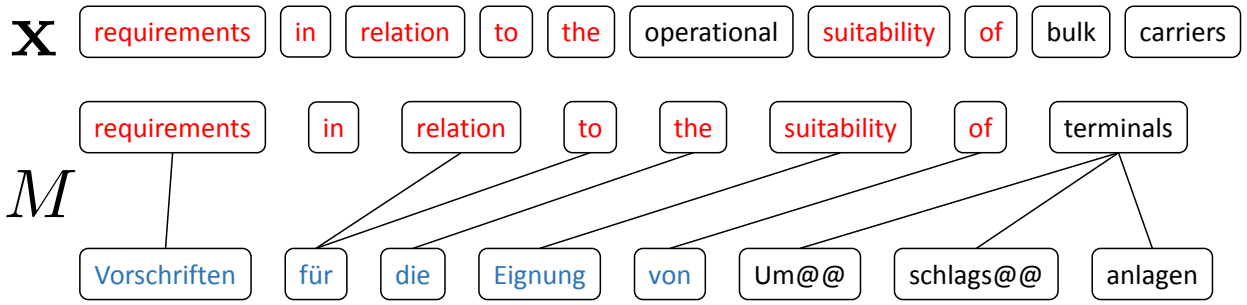


Fig credit: J. Zhang, M. Utiyama, E. Sumita, G. Neubig, S. Nakamura. Guiding Neural Machine Translation with Retrieved Translation Pieces. NAACL18.

Dual model

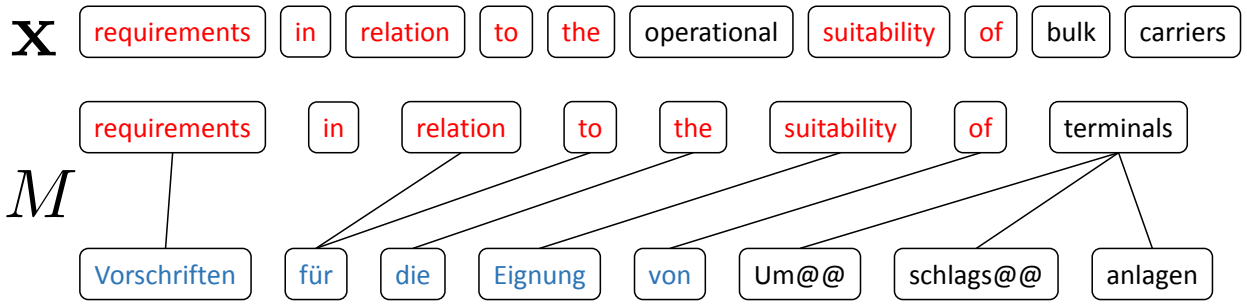


Weighted n-gram

(Vorschriften, 0.8)	(Vorschriften fur, 0.8)
(fur, 0.8)	(fur die, 0.8)
(die, 0.8)	… …
(Eignung, 0.8)	(Vorschriften fur die Eignung, 0.8)
(von, 0.8)	(fur die Eignung von, 0.8)

Fig credit: J. Zhang, M. Utiyama, E. Sumita, G. Neubig, S. Nakamura. Guiding Neural Machine Translation with Retrieved Translation Pieces. NAACL18.

Dual model



Weighted n-gram

(Vorschriften, 0.8)	(Vorschriften fur, 0.8)
(fur, 0.8)	(fur die, 0.8)
(die, 0.8)	… …
(Eignung, 0.8)	(Vorschriften fur die Eignung, 0.8)
(von, 0.8)	(fur die Eignung von, 0.8)

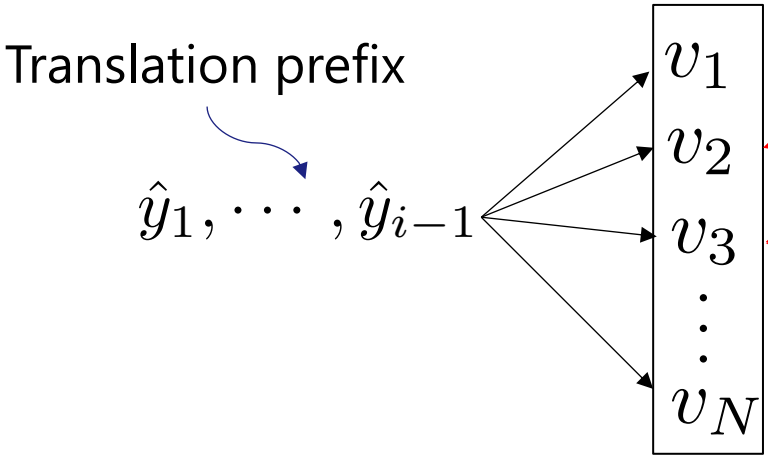
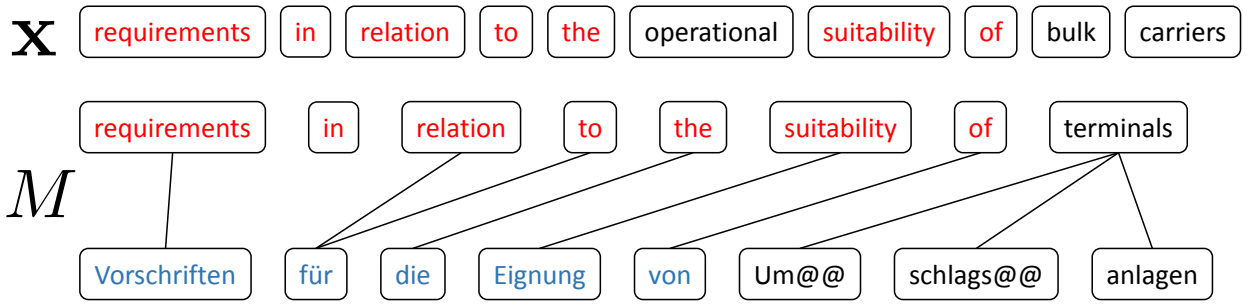


Fig credit: J. Zhang, M. Utiyama, E. Sumita, G. Neubig, S. Nakamura. Guiding Neural Machine Translation with Retrieved Translation Pieces. NAACL18.

Dual model



Weighted n-gram

(Vorschriften, 0.8)
(fur, 0.8)
(die, 0.8)
(Eignung, 0.8)
(von, 0.8)
(Vorschriften fur, 0.8)
(fur die, 0.8)
… …
(Vorschriften fur die Eignung, 0.8)
(fur die Eignung von, 0.8)

Matched n-gram

Translation prefix

$\hat{y}_1, \cdots, \hat{y}_{i-1}$

v_1
v_2
v_3
$\vdots$
v_N

$p_{\mathrm{TM}}(v_2)$
$p_{\mathrm{TM}}(v_3)$

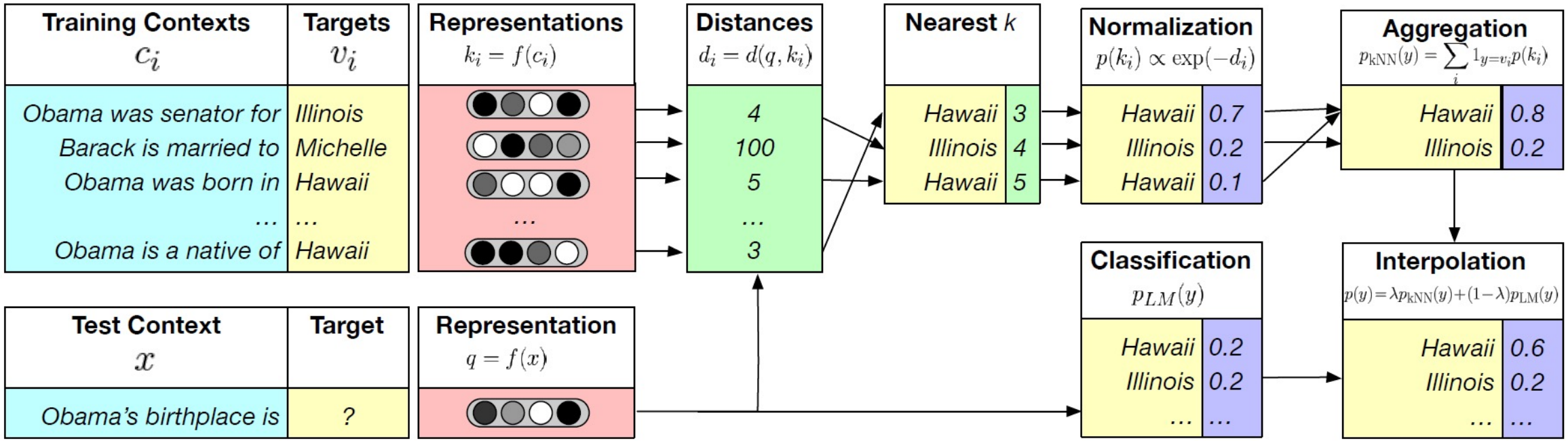
v_2
v_3
$\hat{y}_{i-1}, v_3$
$\hat{y}_{i-2}, \hat{y}_{i-1}, v_3$

Standard NMT model
(RNN, Transformer)

$$p(y_i|\mathbf{x}, \hat{\mathbf{y}}_{1:i-1}) = p_{\mathrm{NMT}}(y_i|\mathbf{x}, \hat{\mathbf{y}}_{1:i-1}) + \lambda \times p_{\mathrm{TM}}(y_i)$$

Fig credit: J. Zhang, M. Utiyama, E. Sumita, G. Neubig, S. Nakamura. Guiding Neural Machine Translation with Retrieved Translation Pieces. NAACL18.

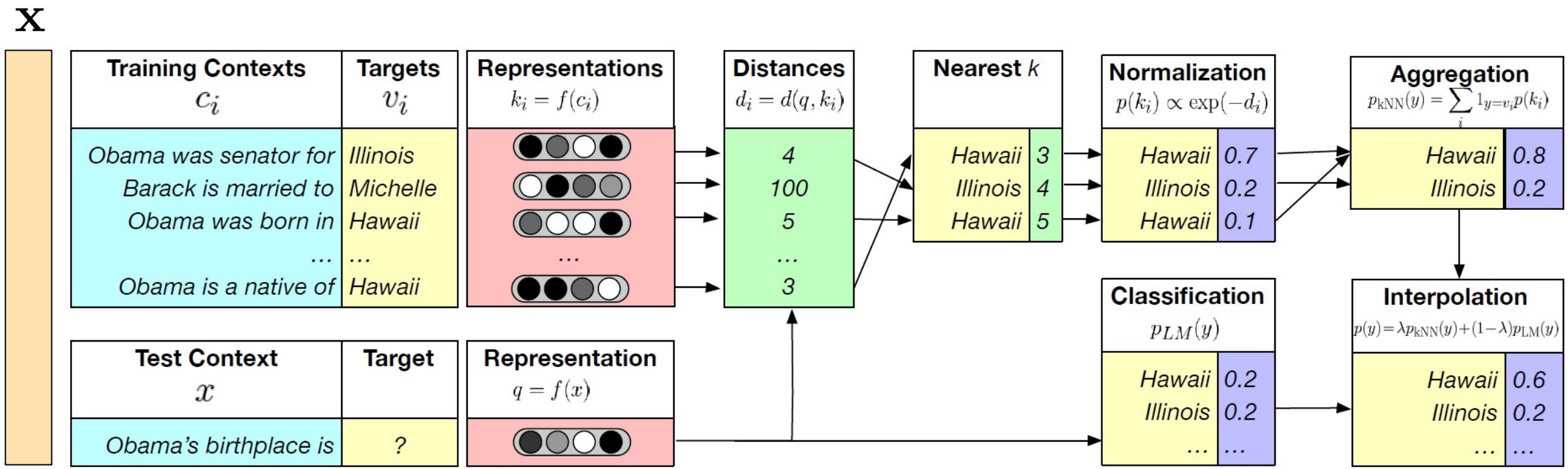
Dual model: KNN-MT Extended from KNN-LM



KNN-LM

Fig. Credit: U. Khandelwal, O. Levy, D. Jurafsky, L. Zettlemoyer, M. Lewis. Generalization through Memorization: Nearest Neighbor Language Models. ICLR 20.

Dual model: KNN-MT Extended from KNN-LM



KNN-LM

Fig. Credit: U. Khandelwal, O. Levy, D. Jurafsky, L. Zettlemoyer, M. Lewis. Generalization through Memorization: Nearest Neighbor Language Models. ICLR 20.

Dual model: KNN-MT

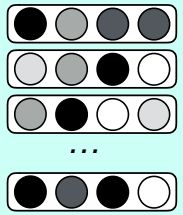
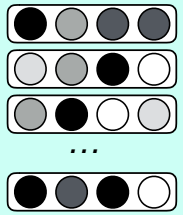
Training Translation Contexts $(s^{(n)}, t_{i-1}^{(n)})$		Datastore	
		Representation $k_j = f(s^{(n)}, t_{i-1}^{(n)})$	Target $v_j = t_i^{(n)}$
J'ai été à Paris.	*I have*		*been*
J'avais été à la maison.	*I had*		*been*
J'apprécie l'été.	*I enjoy*		*summer*
…	…	…	…
J'ai ma propre chambre.	*I have*		*my*

Fig. Credit: Urvashi Khandelwal, Angela Fan, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. Nearest neighbor machine translation. ICLR21.

Dual model: KNN-MT

Training Translation Contexts $(s^{(n)}, t_{i-1}^{(n)})$		Datastore	
		Representation $k_j = f(s^{(n)}, t_{i-1}^{(n)})$	Target $v_j = t_i^{(n)}$
J'ai été à Paris. *J'avais été à la maison.* *J'apprécie l'été.* *…* *J'ai ma propre chambre.*	*I have* *I had* *I enjoy* *…* *I have*		*been* *been* *summer* *…* *my*

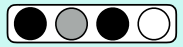
Test Input x	Generated tokens $\hat{y}_{1:i-1}$	Representation $q = f(x, \hat{y}_{1:i-1})$	Target y_i
J'ai été dans ma propre chambre.	*I have*		*?*

Fig. Credit: Urvashi Khandelwal, Angela Fan, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. Nearest neighbor machine translation. ICLR21.

Dual model: KNN-MT

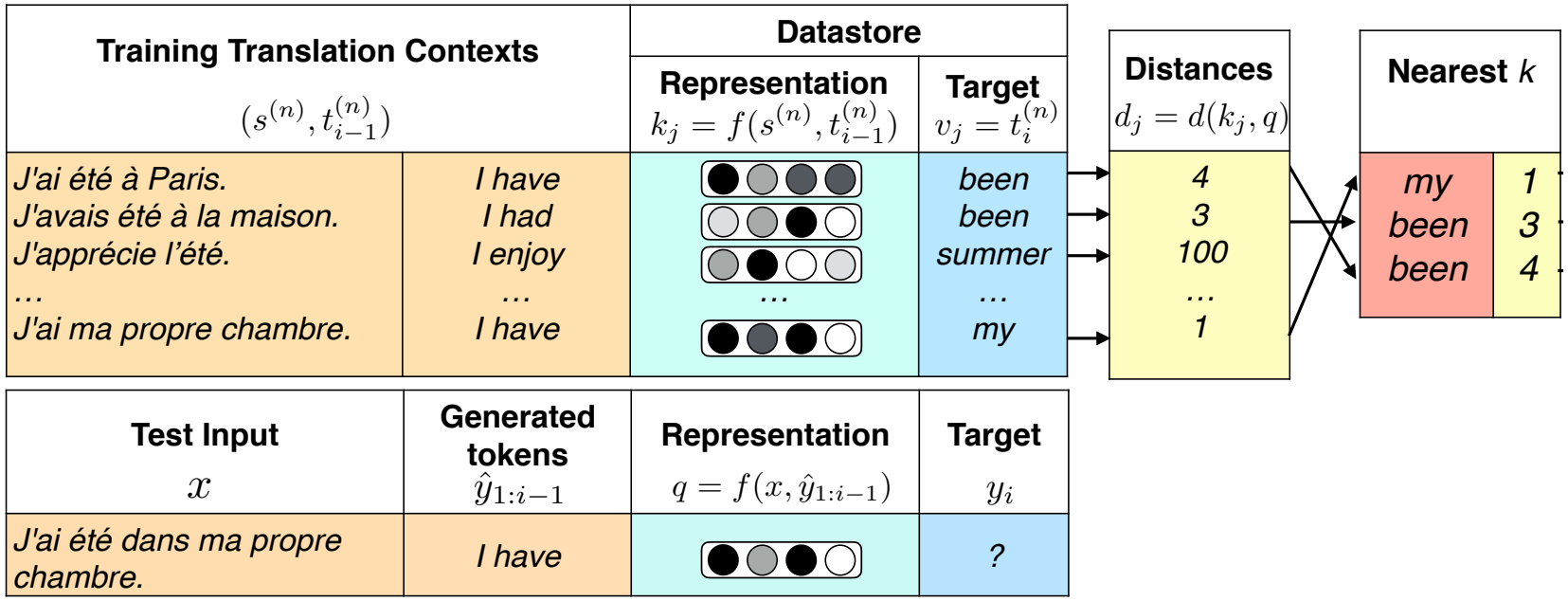


Fig. Credit: Urvashi Khandelwal, Angela Fan, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. Nearest neighbor machine translation. ICLR21.

Dual model: KNN-MT

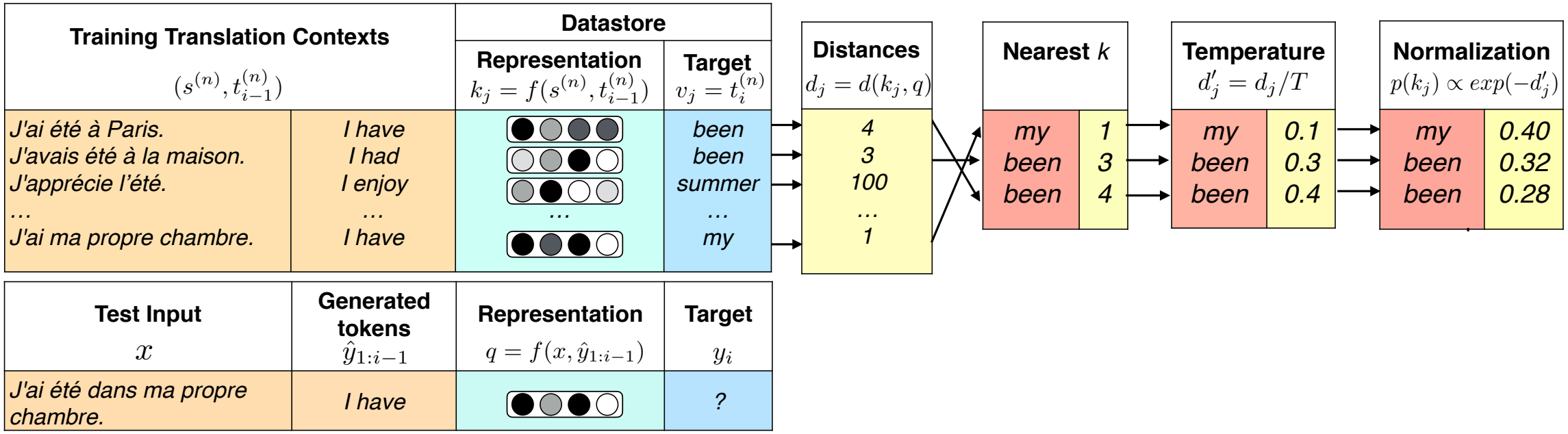


Fig. Credit: Urvashi Khandelwal, Angela Fan, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. Nearest neighbor machine translation. ICLR21.

Dual model: KNN-MT

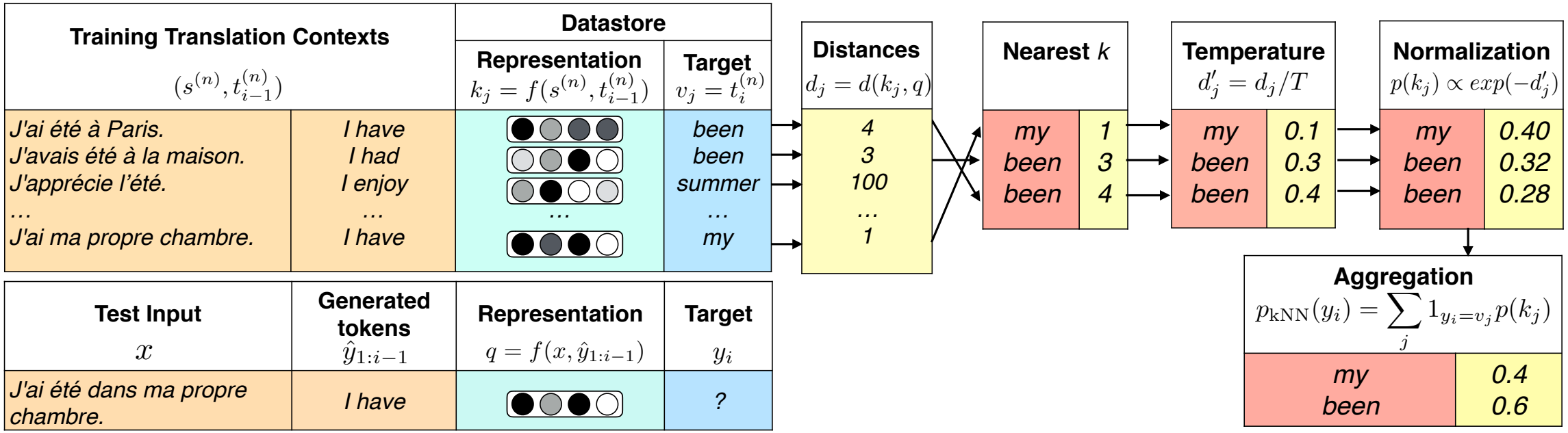
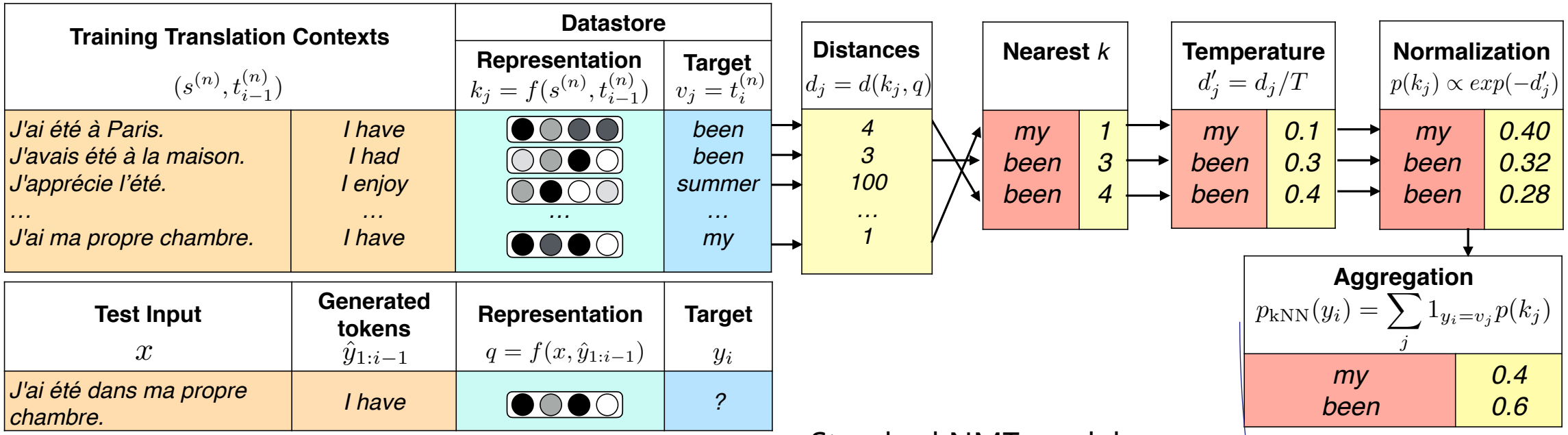


Fig. Credit: Urvashi Khandelwal, Angela Fan, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. Nearest neighbor machine translation. ICLR21.

Dual model: KNN-MT

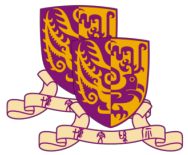


Standard NMT model (RNN, Transformer)

$$p(y_i|\mathbf{x}, \hat{\mathbf{y}}_{1:i-1}) = p_{\text{NMT}}(y_i|x, \hat{\mathbf{y}}_{1:i-1}) + \lambda \times p_{k\text{NN}}(y_i)$$

Fig. Credit: Urvashi Khandelwal, Angela Fan, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. Nearest neighbor machine translation. ICLR21.

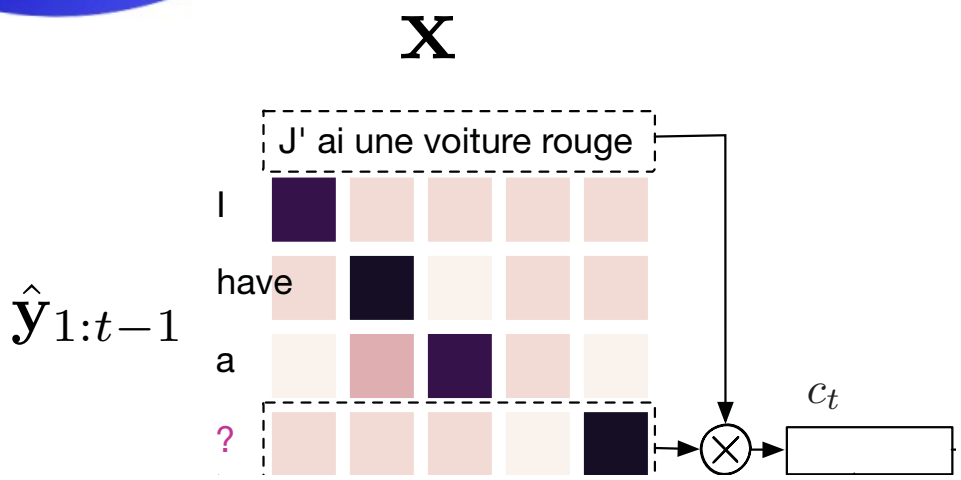
Outline



Tencent
AI Lab

- Background and Introduction
- Language Modeling
- Open-Domain Dialogue Systems
- **Neural Machine Translation**
  - Motivation
  - TM-augmented NMT Framework
  - **TM-augmented Models**
    - Standard model
    - Dual model
    - **Unified model**
- Conclusion and Outlook

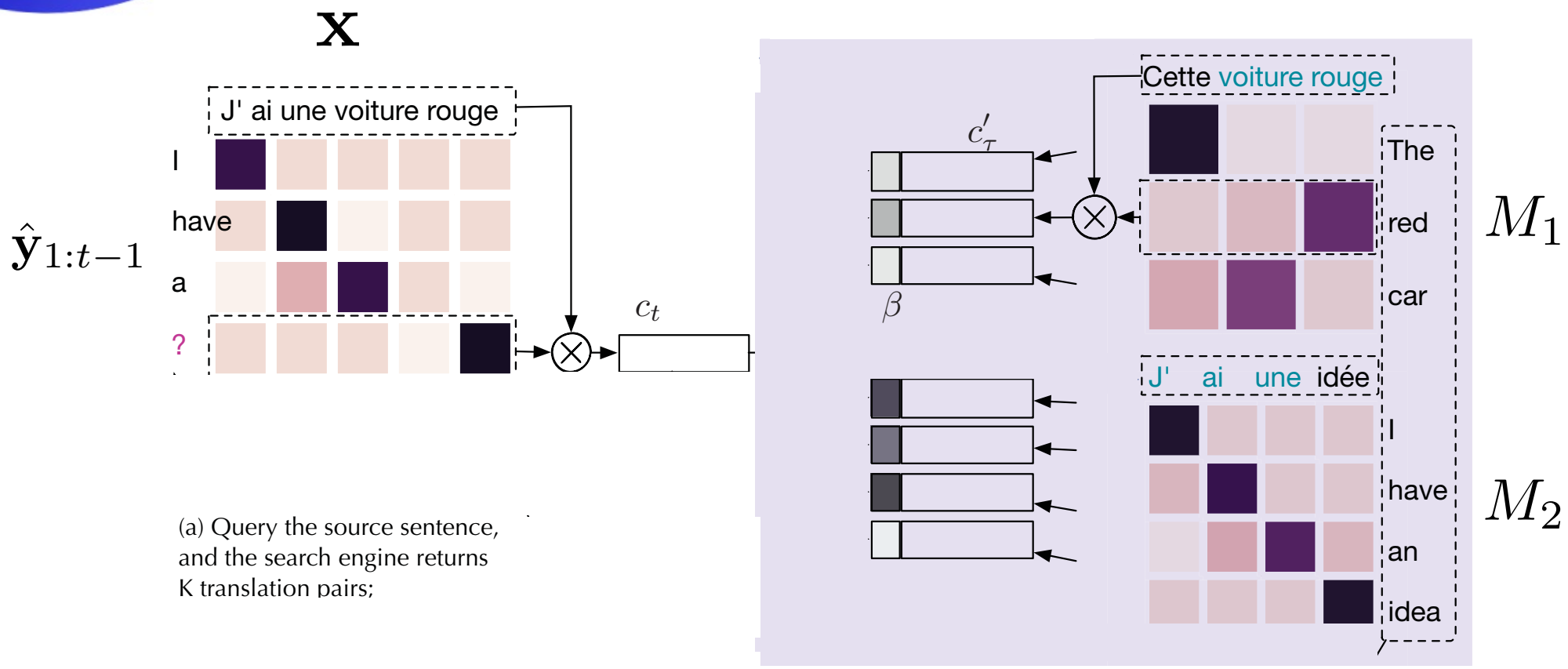
Unified Model: CopyNet on TM



(a) Query the source sentence, and the search engine returns K translation pairs;

Fig. Credit: Jiatao Gu, Yong Wang, Kyunghyun Cho, Victor O.K. Li. Search Engine Guided Neural Machine Translation. AAAI18.

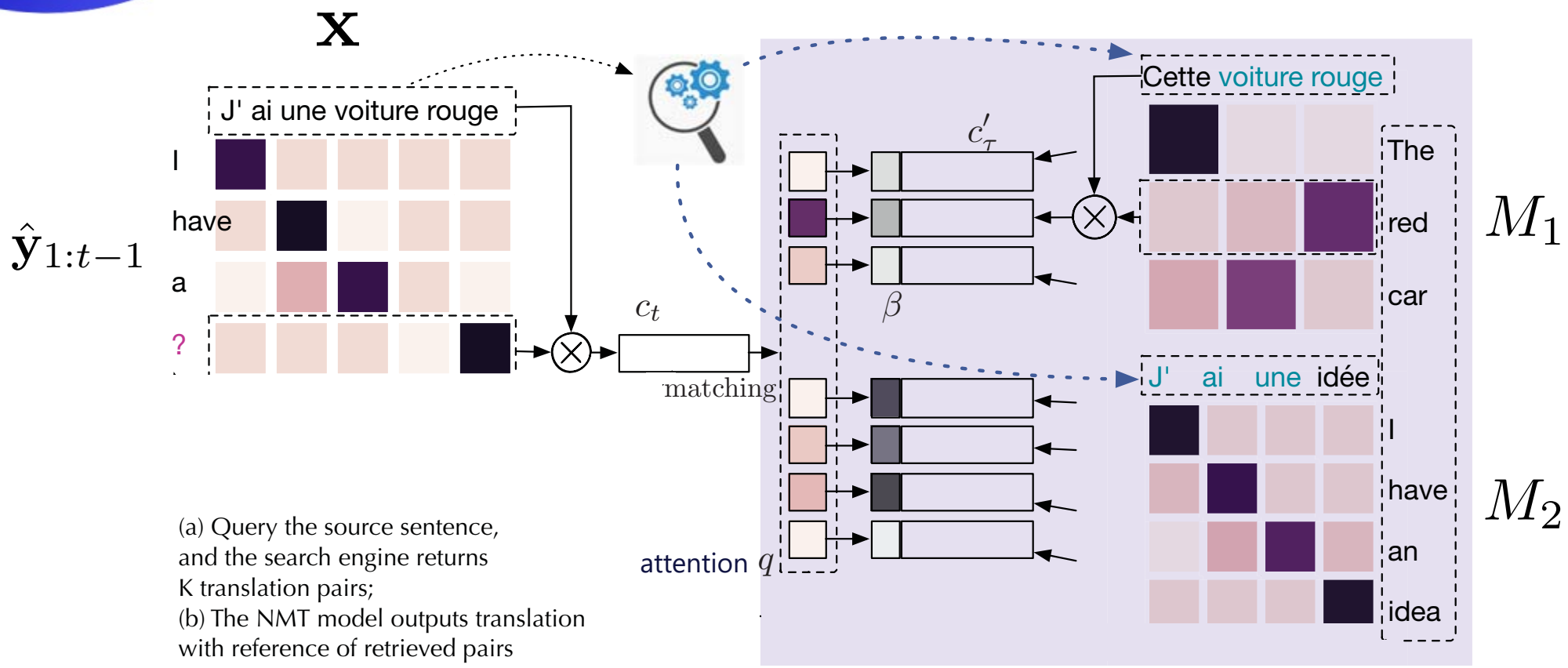
Unified Model: CopyNet on TM



(a) Query the source sentence, and the search engine returns K translation pairs;

Fig. Credit: Jiatao Gu, Yong Wang, Kyunghyun Cho, Victor O.K. Li. Search Engine Guided Neural Machine Translation. AAAI18.

Unified Model: CopyNet on TM



(a) Query the source sentence, and the search engine returns K translation pairs;
(b) The NMT model outputs translation with reference of retrieved pairs

Fig. Credit: Jiatao Gu, Yong Wang, Kyunghyun Cho, Victor O.K. Li. Search Engine Guided Neural Machine Translation. AAAI18.

Unified Model: CopyNet on TM

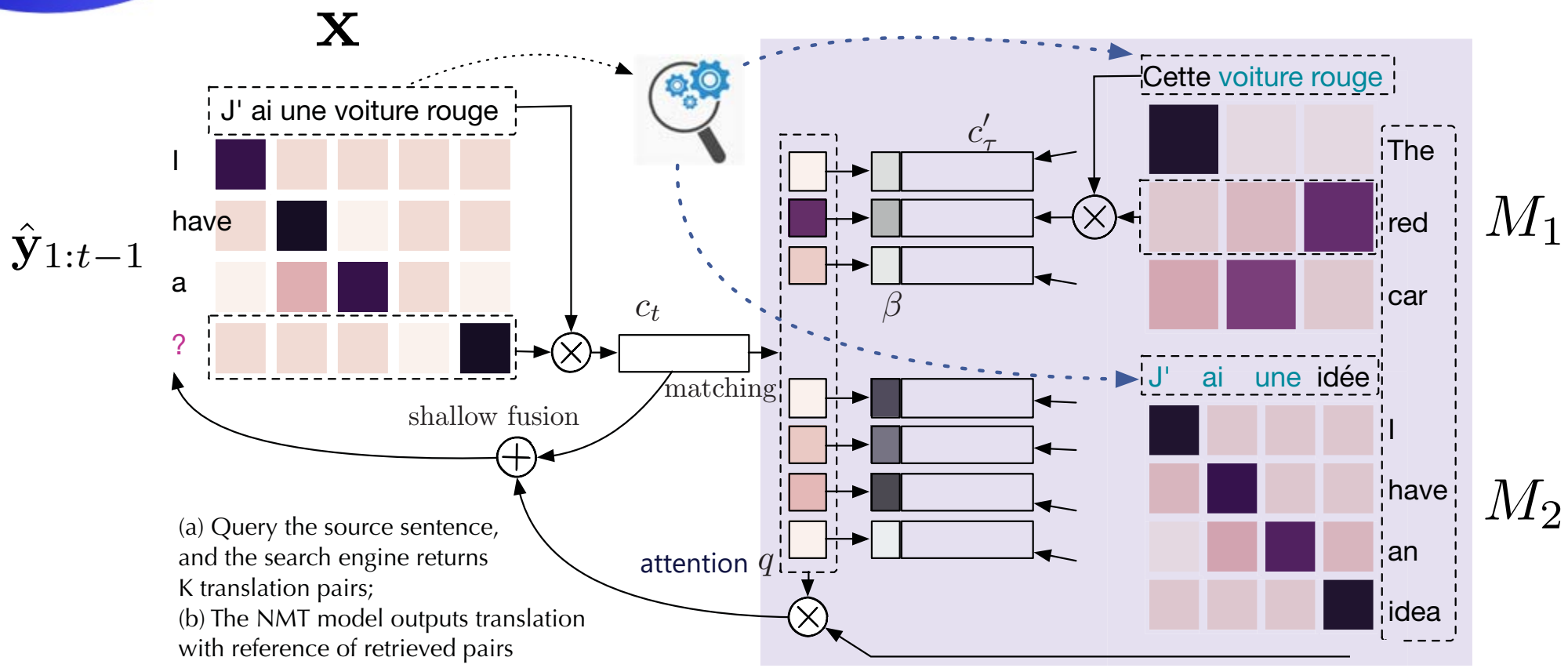
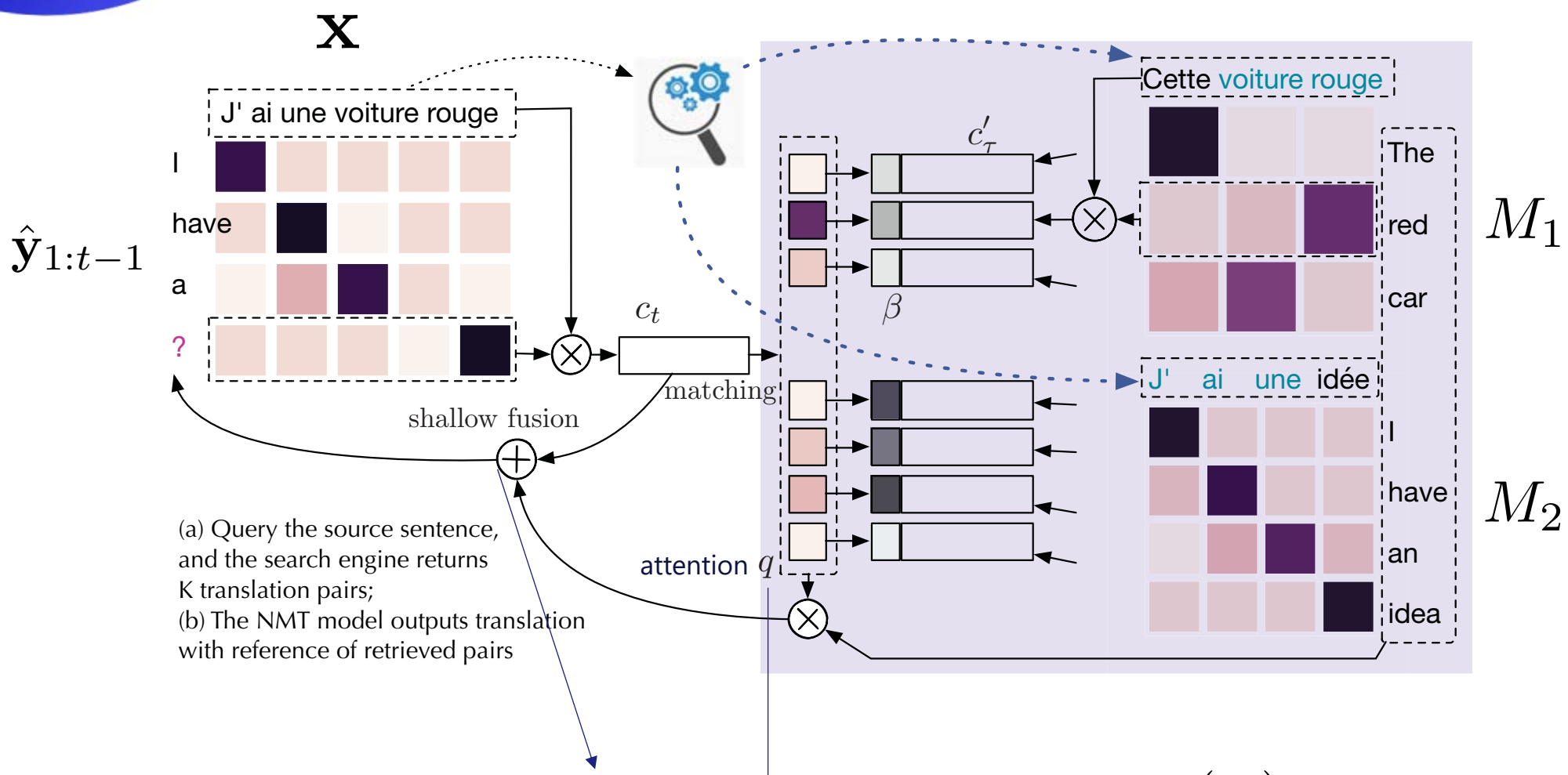


Fig. Credit: Jiatao Gu, Yong Wang, Kyunghyun Cho, Victor O.K. Li. Search Engine Guided Neural Machine Translation. AAAI18.

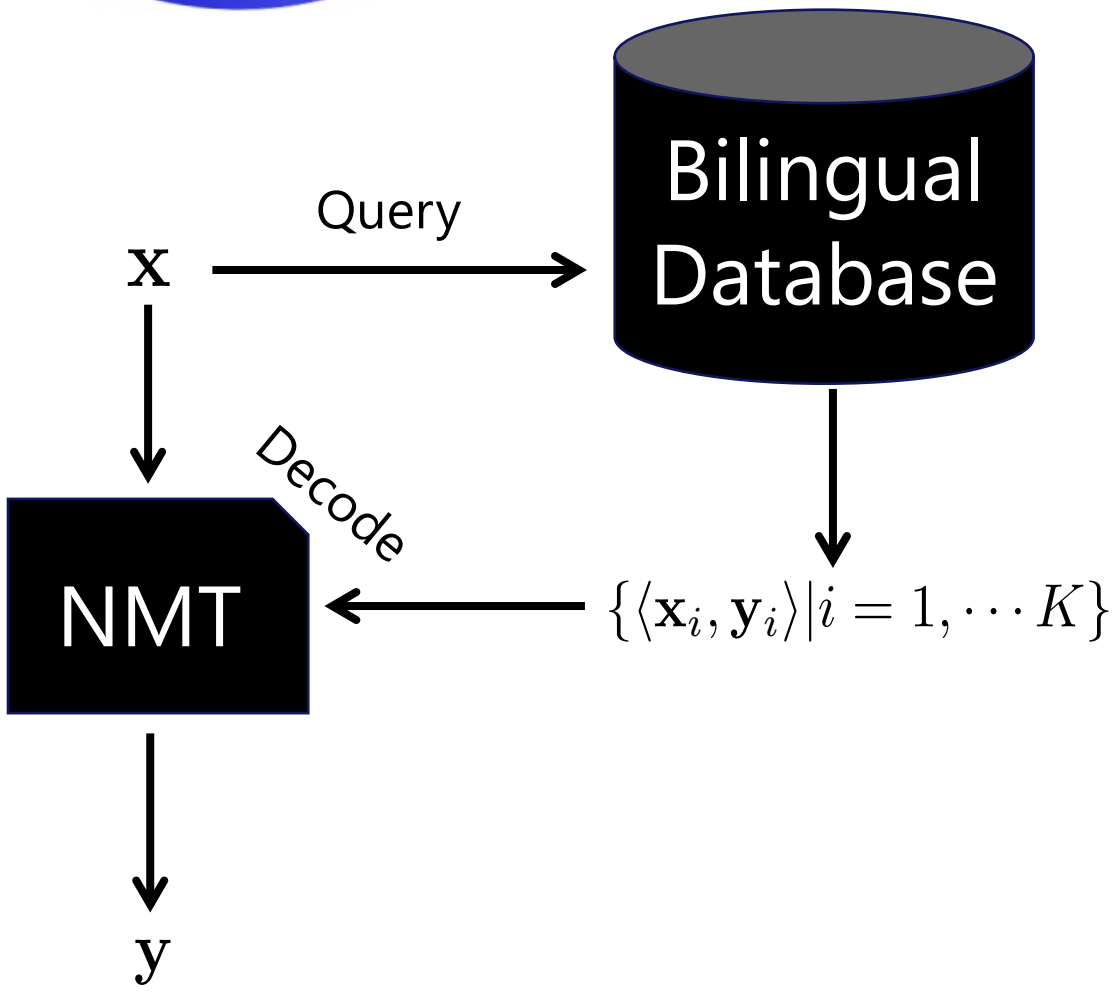
Unified Model: CopyNet on TM



$$p(y_t|\mathbf{x}, \hat{\mathbf{y}}_{1:t-1}, M; \theta) = \zeta_t(\theta) \times p_{\text{copy}}(y_t; \theta) + (1 - \zeta_t(\theta)) p(y_t|\mathbf{x}, \hat{\mathbf{y}}_{1:t-1}; \theta)$$

Fig. Credit: Jiatao Gu, Yong Wang, Kyunghyun Cho, Victor O.K. Li. Search Engine Guided Neural Machine Translation. AAAI18.

Limitations in conventional TM framework

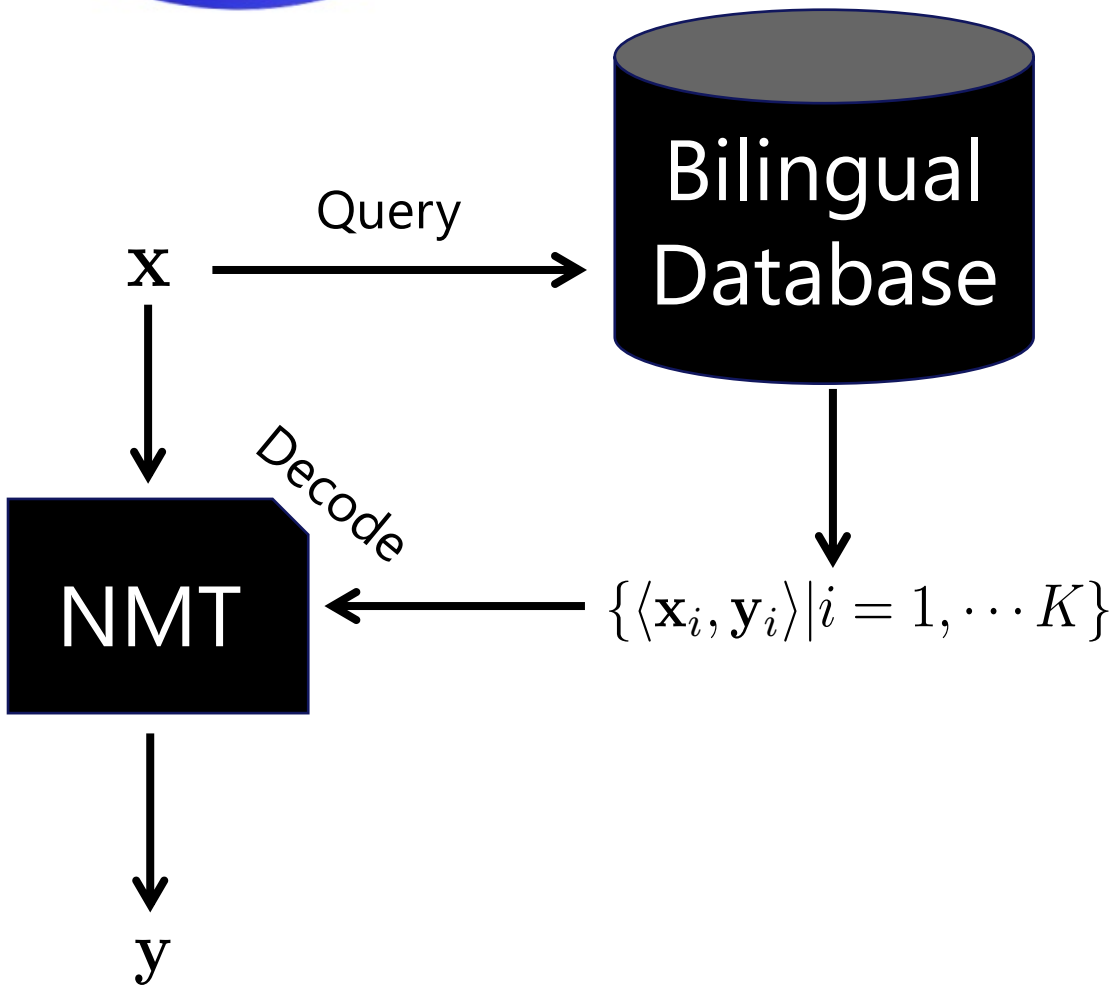


Heuristic retrieval metric

- edit-dist on source sentences
- unaware of translation tasks

Deng Cai, Yan Wang, Huayang Li, Wai Lam, Lemao Liu. Neural Machine Translation with Monolingual Translation Memory. ACL21.

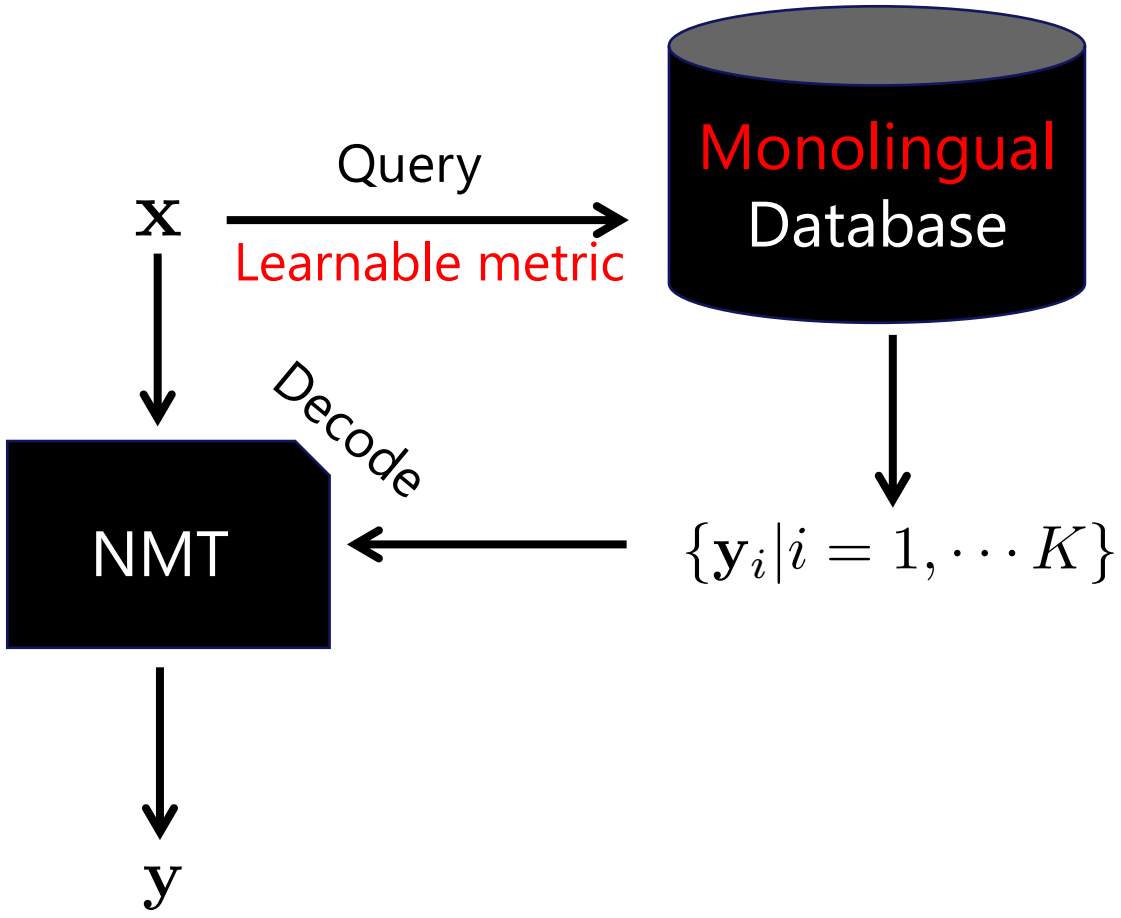
Limitations in conventional TM framework



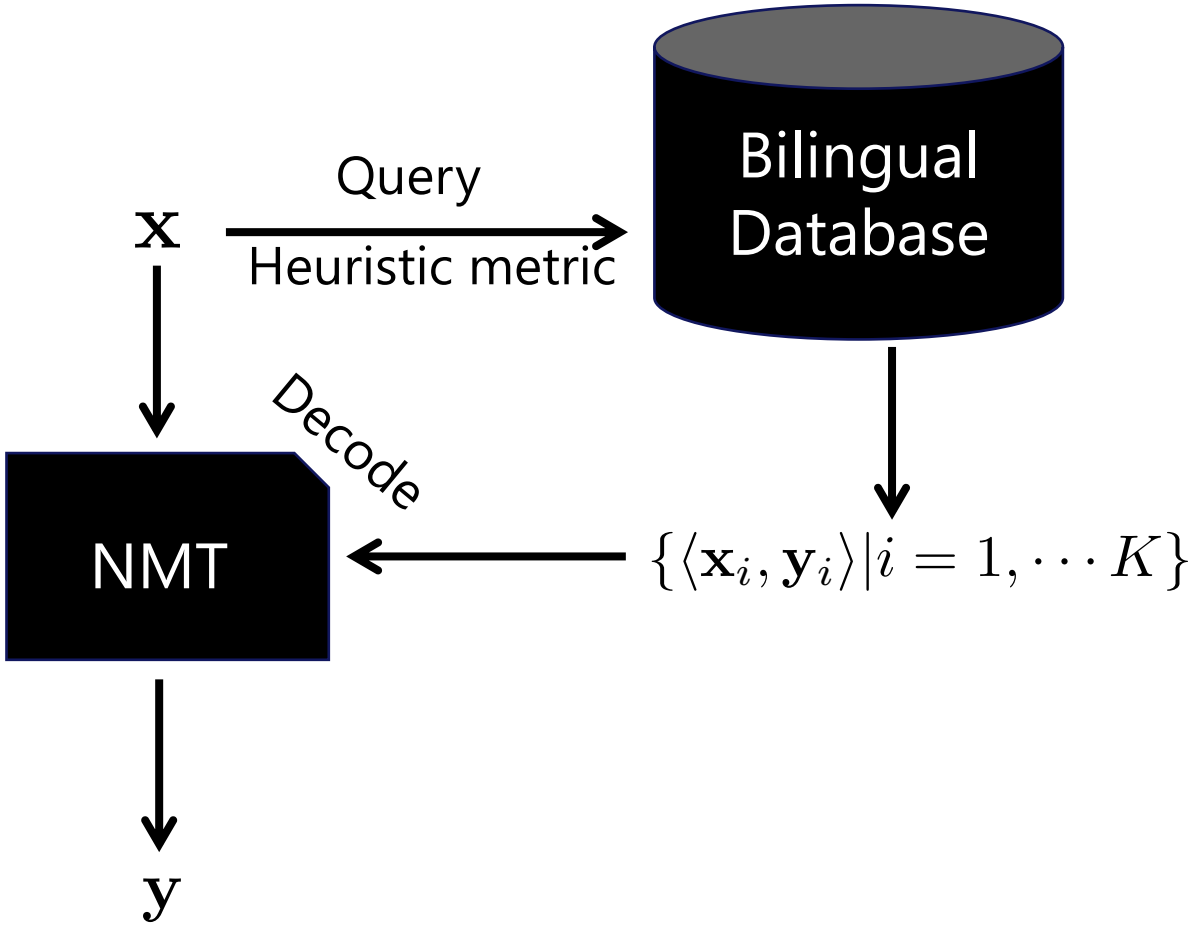
- Limiting Applications
  - Search over bilingual database
  - Bilingual data is too expensive

Deng Cai, Yan Wang, Huayang Li, Wai Lam, Lemao Liu. Neural Machine Translation with Monolingual Translation Memory. ACL21.

Monolingual translation memory



The New Framework



Conventional Framework

Deng Cai, Yan Wang, Huayang Li, Wai Lam, Lemao Liu. Neural Machine Translation with Monolingual Translation Memory. ACL21.

Joint learning retrieval and translation models

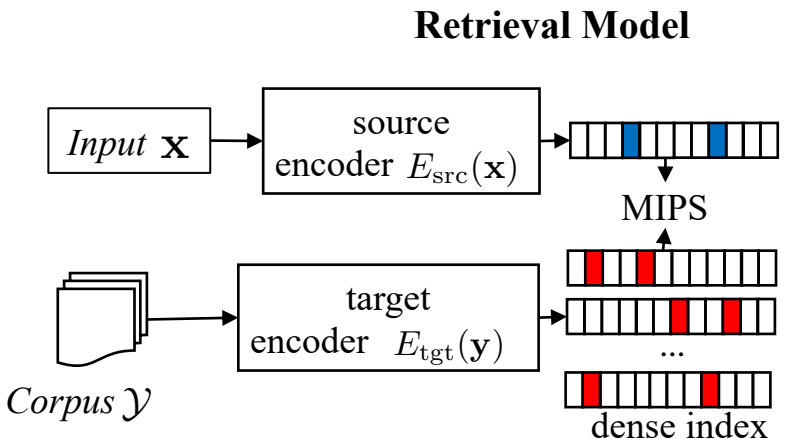
Retrieval Model

Input $\mathbf{x}$

Corpus $\mathcal{Y}$

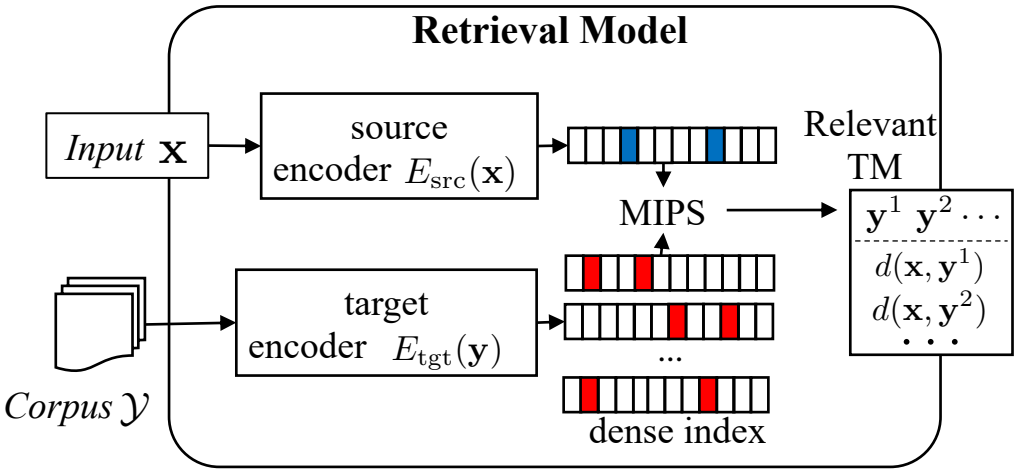
Deng Cai, Yan Wang, Huayang Li, Wai Lam, Lemao Liu. Neural Machine Translation with Monolingual Translation Memory. ACL21.

Joint learning retrieval and translation models



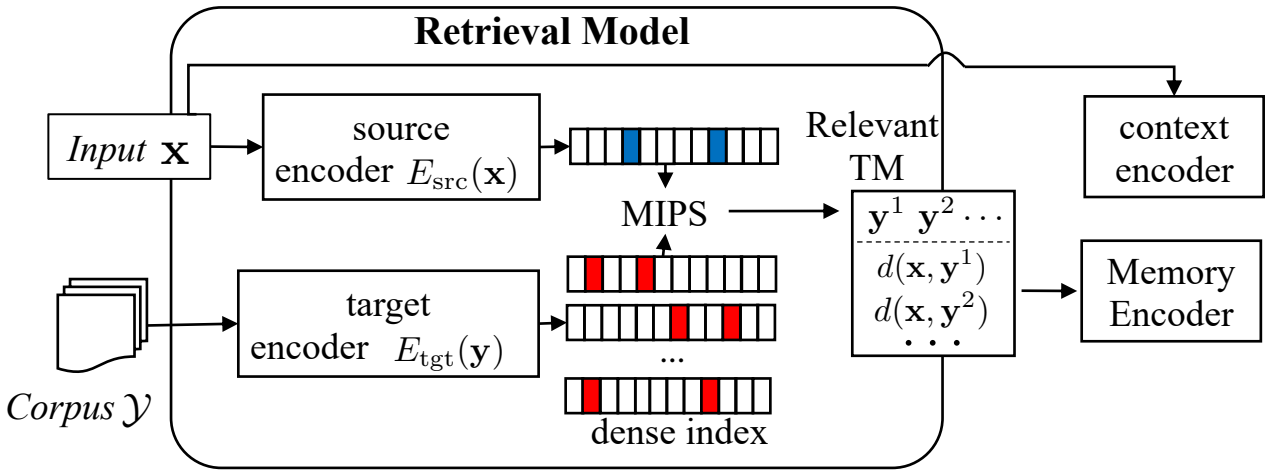
Deng Cai, Yan Wang, Huayang Li, Wai Lam, Lemao Liu. Neural Machine Translation with Monolingual Translation Memory. ACL21.

Joint learning retrieval and translation models



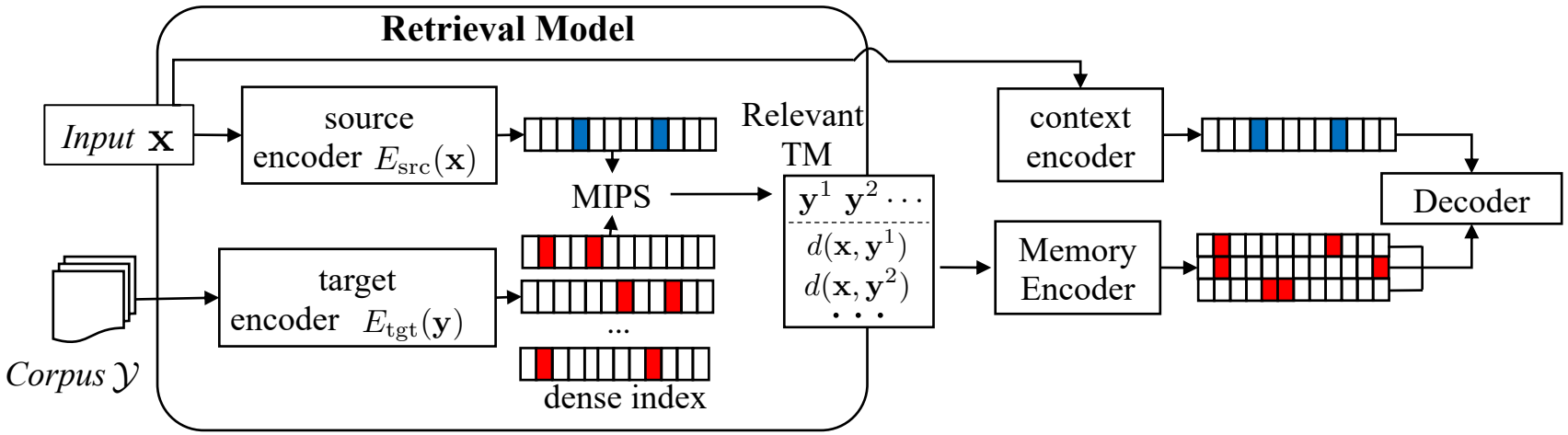
Deng Cai, Yan Wang, Huayang Li, Wai Lam, Lemao Liu. Neural Machine Translation with Monolingual Translation Memory. ACL21.

Joint learning retrieval and translation models



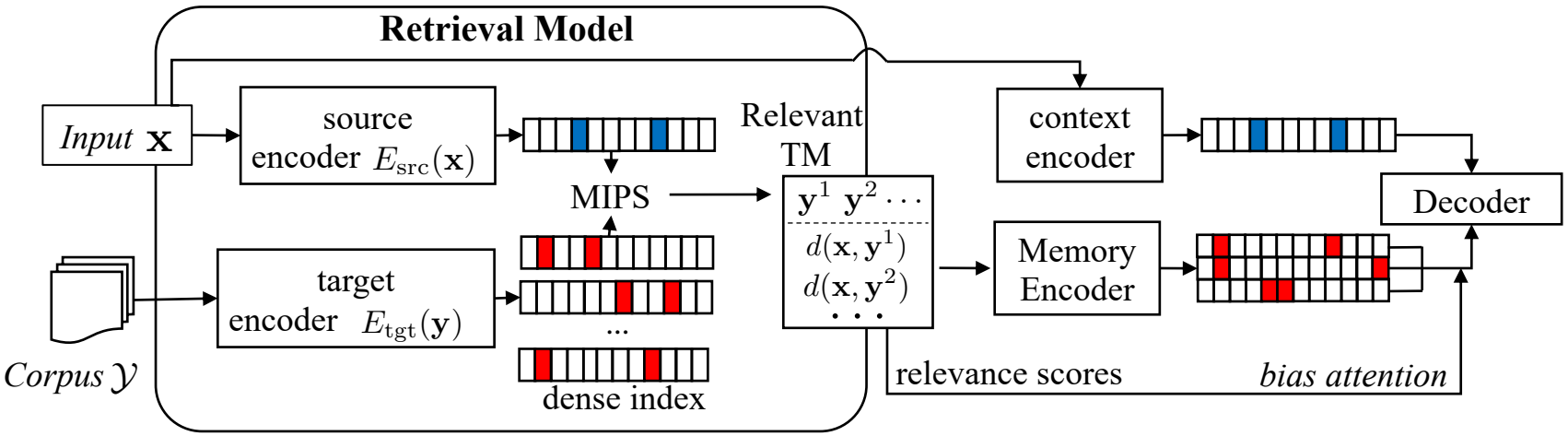
Deng Cai, Yan Wang, Huayang Li, Wai Lam, Lemao Liu. Neural Machine Translation with Monolingual Translation Memory. ACL21.

Joint learning retrieval and translation models



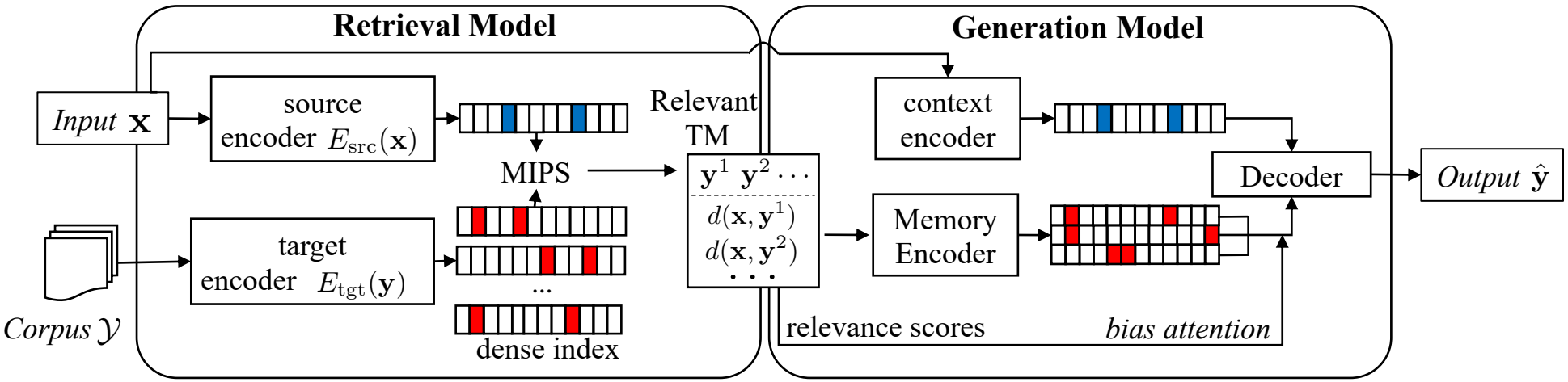
Deng Cai, Yan Wang, Huayang Li, Wai Lam, Lemao Liu. Neural Machine Translation with Monolingual Translation Memory. ACL21.

Joint learning retrieval and translation models



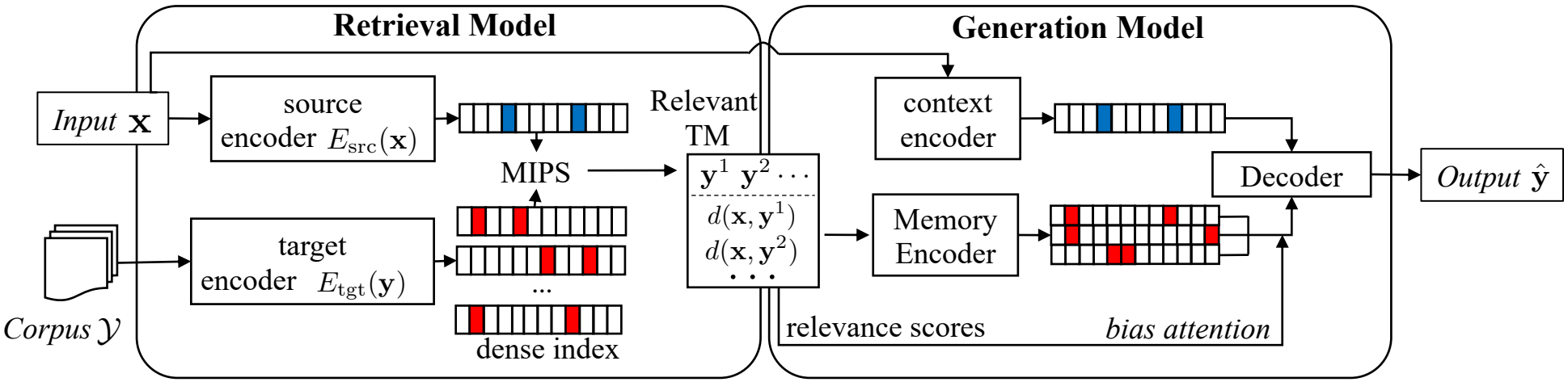
Deng Cai, Yan Wang, Huayang Li, Wai Lam, Lemao Liu. Neural Machine Translation with Monolingual Translation Memory. ACL21.

Joint learning retrieval and translation models



Deng Cai, Yan Wang, Huayang Li, Wai Lam, Lemao Liu. Neural Machine Translation with Monolingual Translation Memory. ACL21.

Joint learning retrieval and translation models



- Challenge: standard MLE training leads to a trivial retrieval metric.
  - Solution: two pre-training subtasks as regularization

Deng Cai, Yan Wang, Huayang Li, Wai Lam, Lemao Liu. Neural Machine Translation with Monolingual Translation Memory. ACL21.

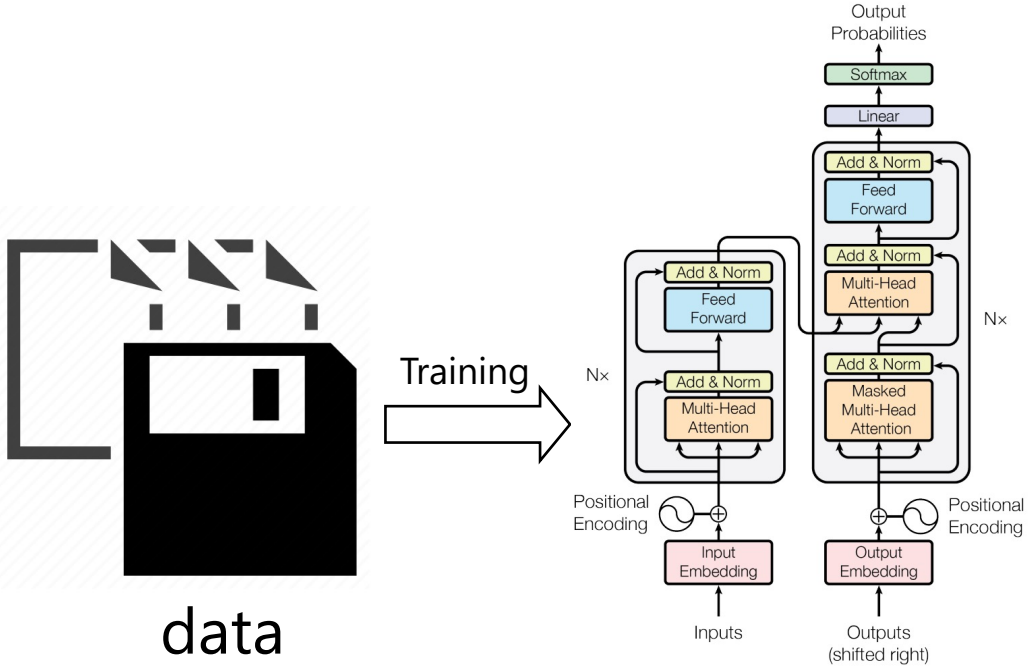
Outline

- Background and Introduction
- Language Modeling
- Open-Domain Dialogue Systems
- Neural Machine Translation
  - Motivation
  - TM-augmented NMT Framework
  - TM-augmented Models
    - Standard model
    - Dual model
    - Unified model
- **Conclusion and Outlook**

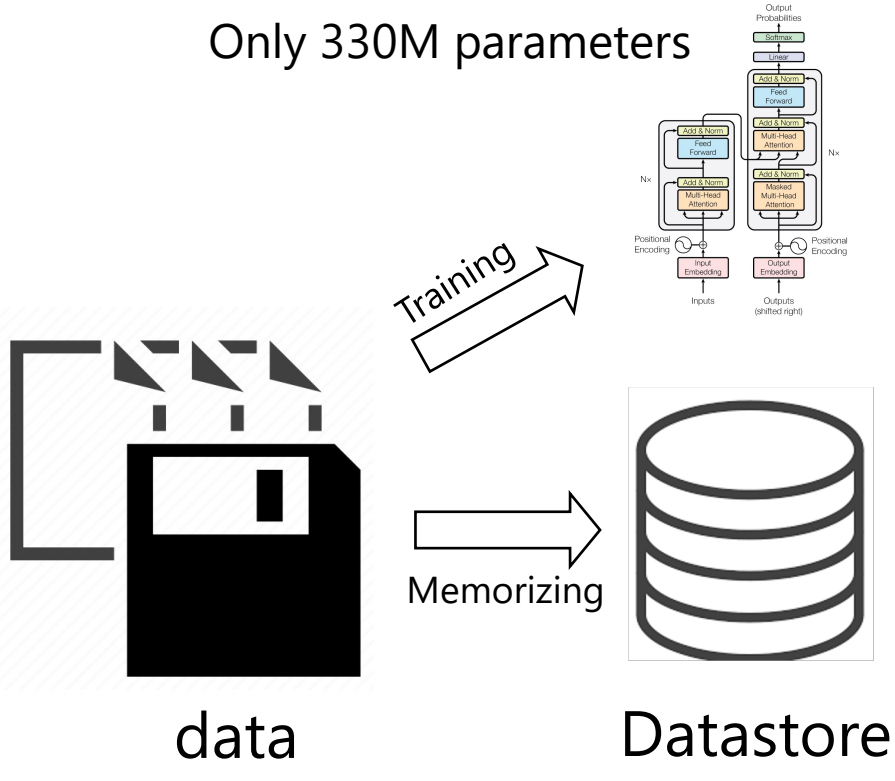
Advantages of retrieval-augmented model

- Compact model with less parameters
  - The knowledge is not implicitly stored in model parameters but in memory

T5 with 11318M parameters



Only 330M parameters



Advantages of retrieval-augmented model

- Better interpretability
  - Some prediction results can be explained through the cues in memory.

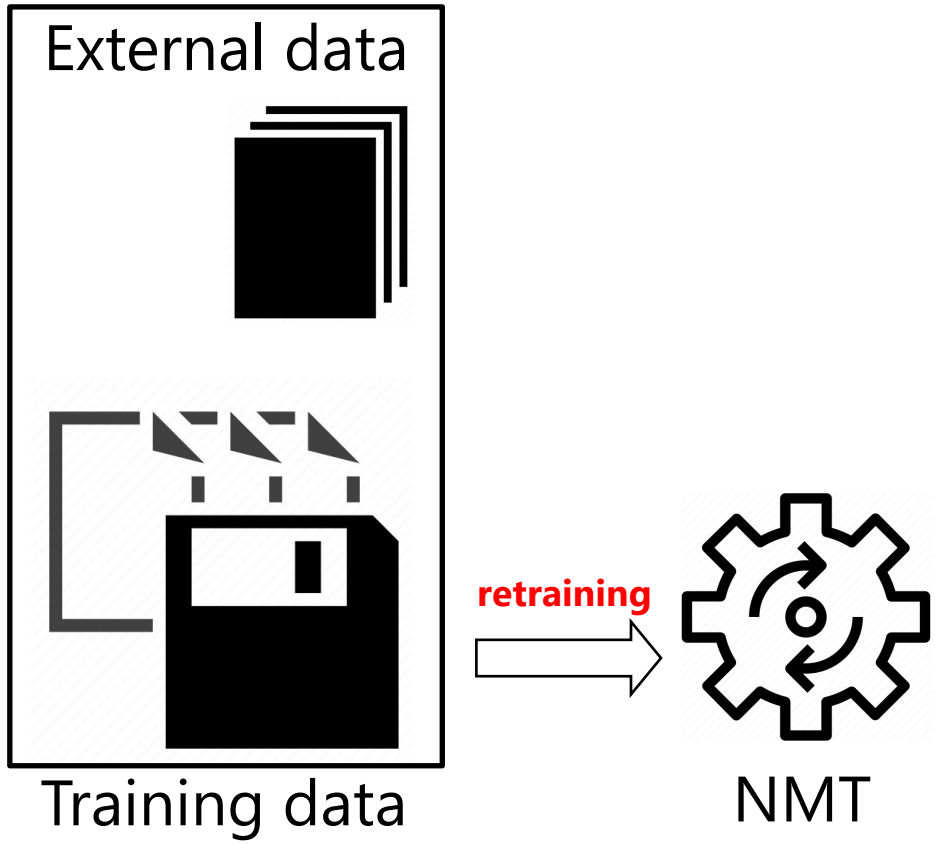
Memory

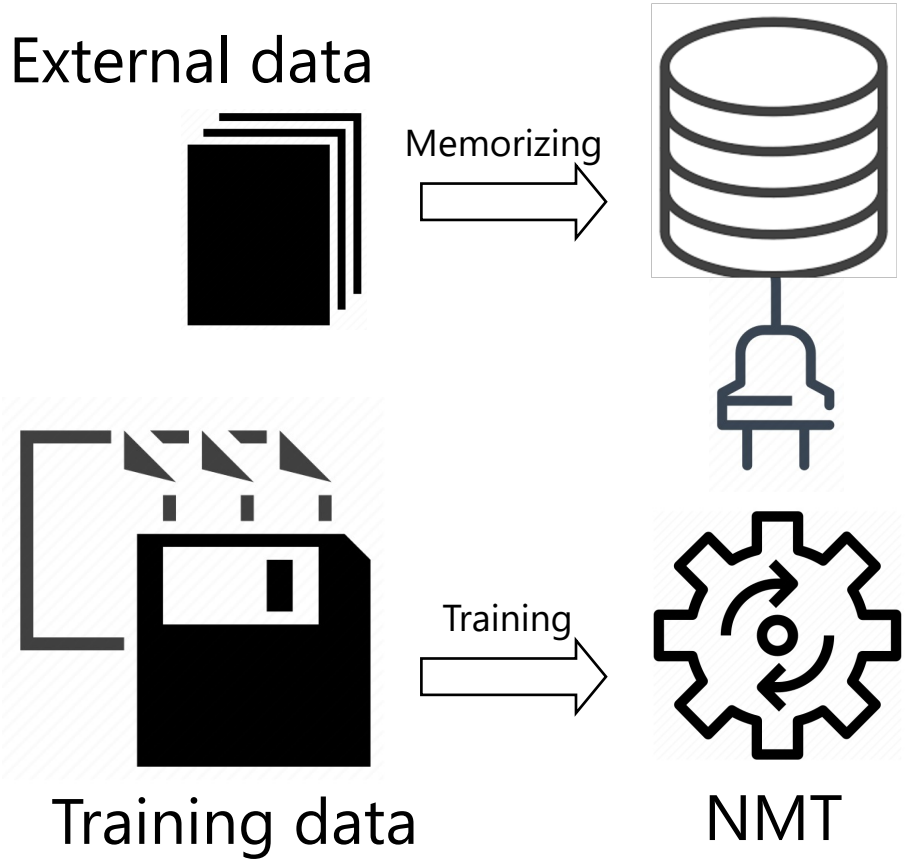
IJCAI 2022 will be held in Vienna , **which is the capital and the largest city of Austria .**

Text Generation by retrieval augmented LM

Advantages of retrieval-augmented model

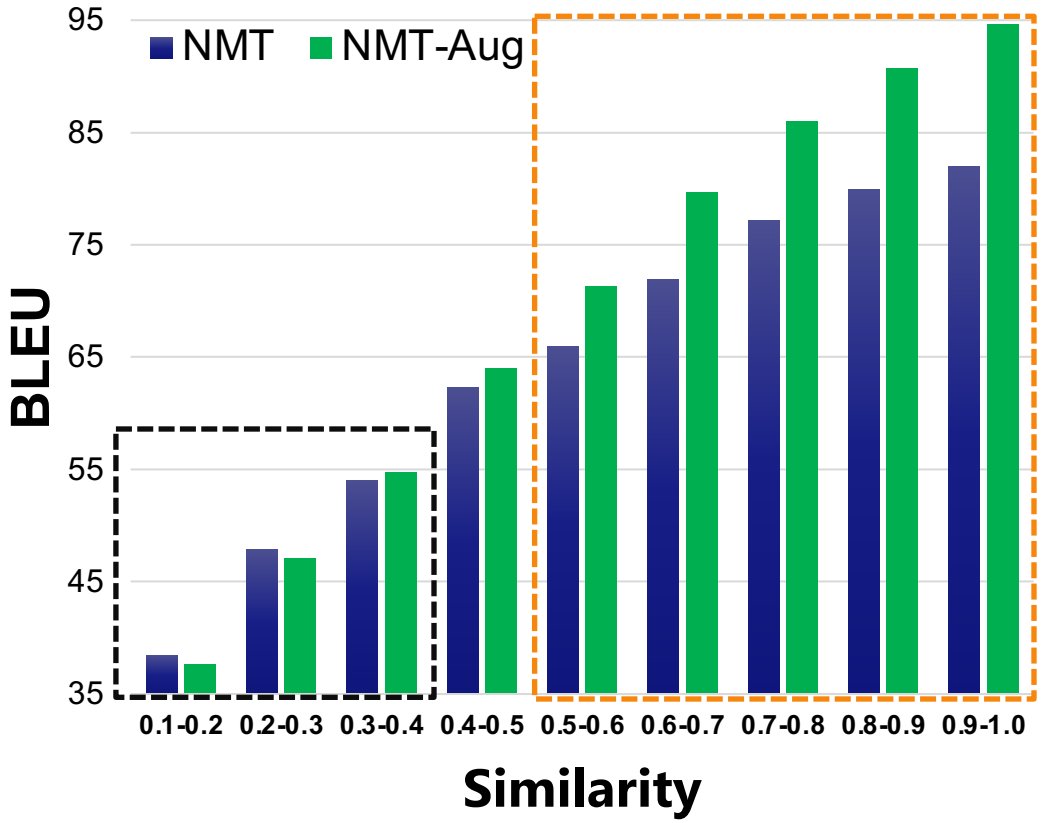
- Better scalability
  - External data can be used as memory in a plug-and-play manner, leading to great scalability





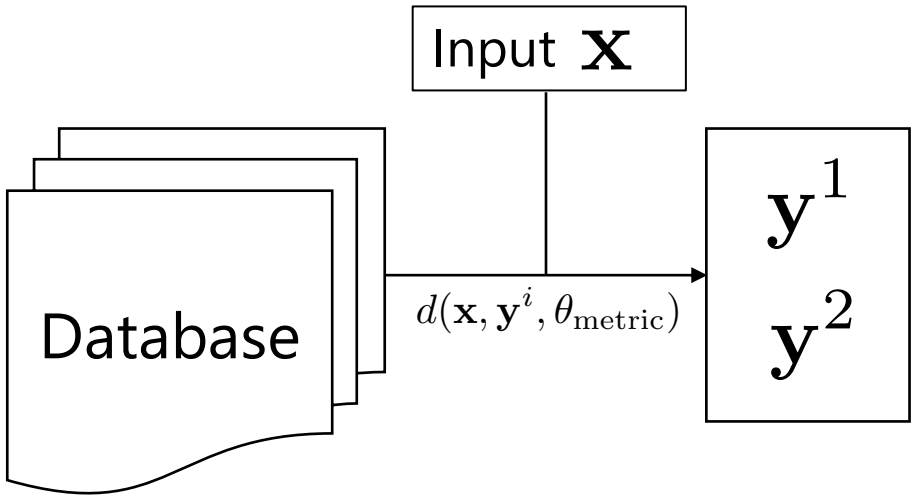
Future Directions

- Retrieval sensitivity
  - Substantial gains for test sentences with high quality memory
  - No gains for those with low quality memory
  - How to alleviate the sensitivity issue?

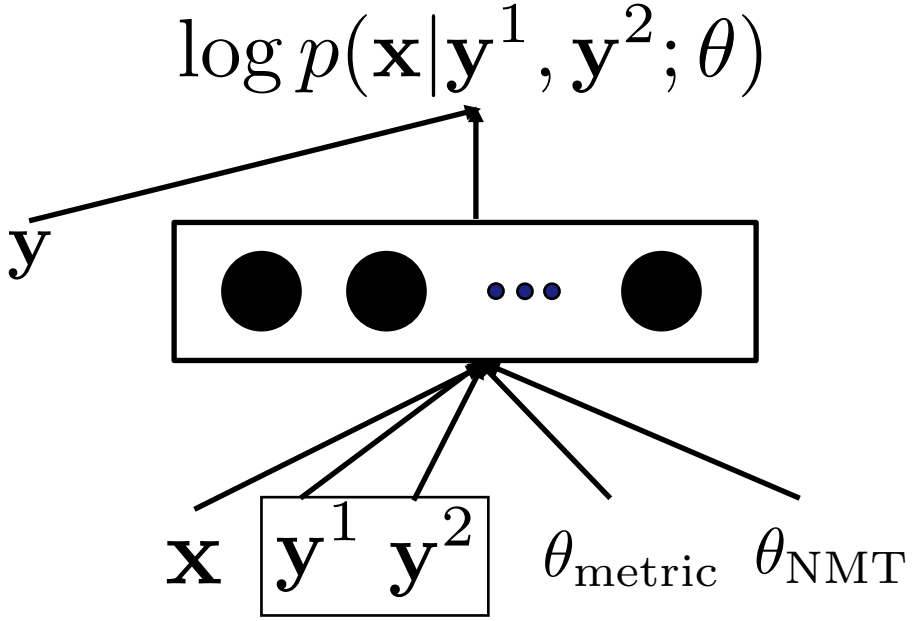


Future Directions

- Gap when jointly learning retrieval metrics towards translation quality
  - Global retrieval: retrieval is conducted in the entire database
  - Local optimization: the parameters are optimized with respect to a tiny fraction of database.



Global Retrieval

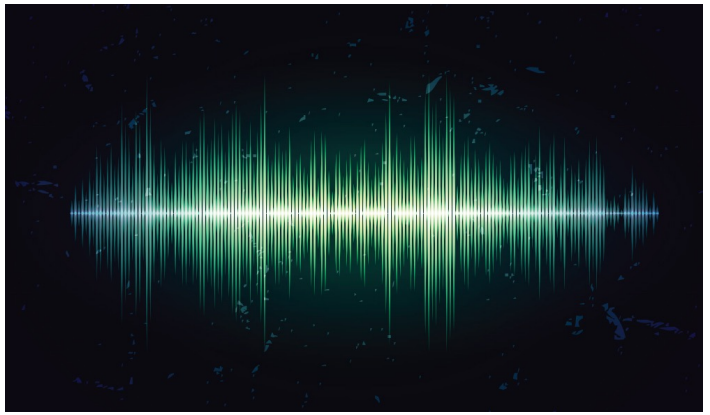


Local optimization

Future Directions

- Retrieval from multi-modality database
  - Most existing works focus on generation models augmented by text memory
  - Multi-modality information can provide complementary information for generation models

Image database



Audio database

Video database

Q&A

Thanks