Evolution Strategy Based Automatic Tuning of Neural Machine Translation Systems

HAO QIN, Takahiro Shinozaki （Tokyo Tech）
Kevin Duh （JHU）

JOHNS HOPKINS UNIVERSITY

Introduction

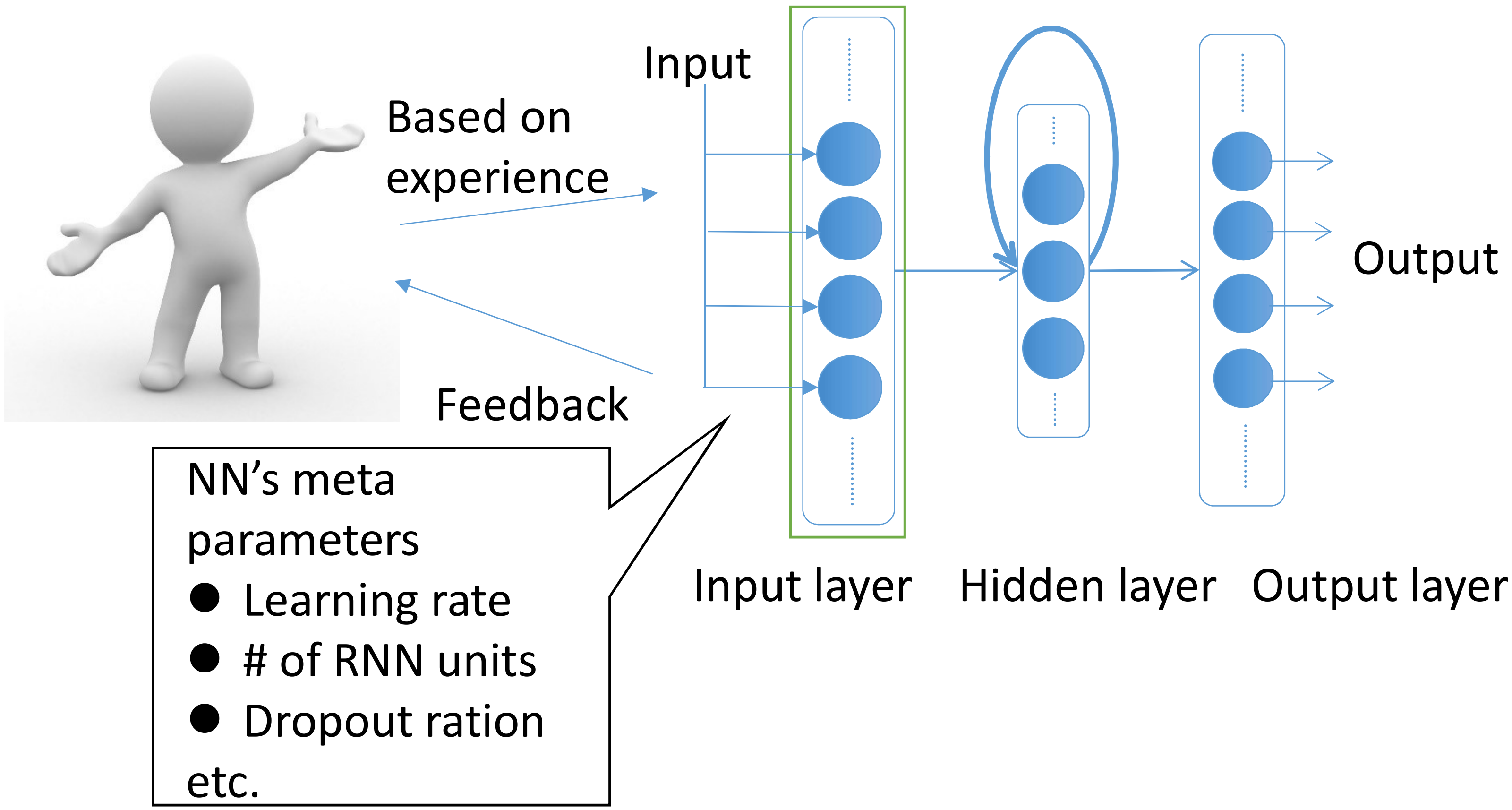
- Neural machine translation (NMT) system have demonstrated promising results in recent years

Google's neural machine translation system reduced translation errors when compared with the prior Google translation technology

The major design question of using neural network structure is how to set the meta-parameter values of the network structure and training configures

Problems of neural network tuning

- Human tuning (“tuning” refers to meta-parameters search)



● Tuning by human experts requires a lot of effort

Related work

- Grid search

  - A simple method for meta-parameter optimization

  -Becomes less tractable as # of parameters increases

- Genetic algorithms(GA), Bayesian optimization(BO)

  -Demonstrated success in many practical problems

- In our work, we apply CMA-ES (covariance matrix adaptive evolutionary strategy) to NMT

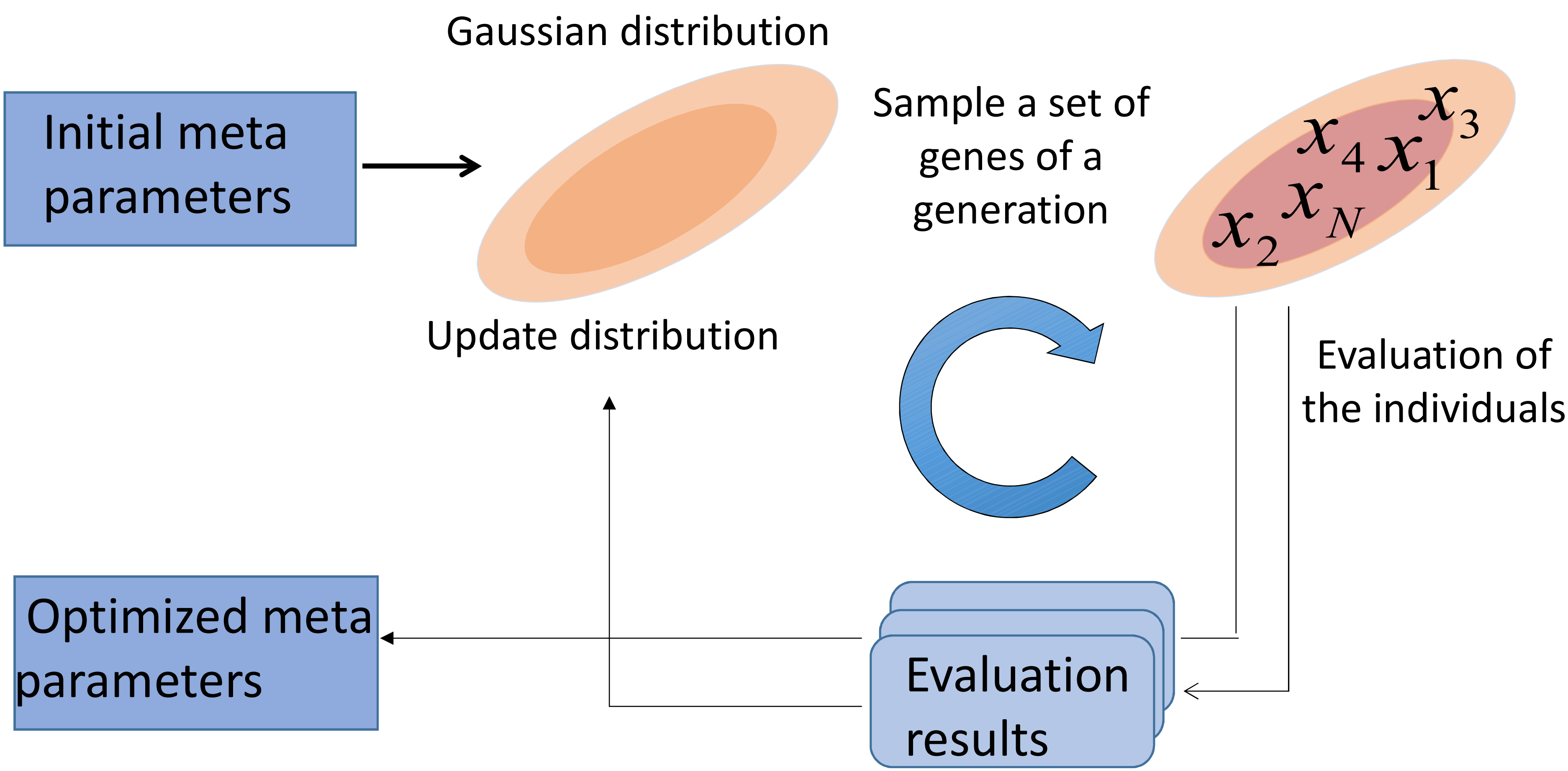
-Previous work shows CMA-ES works to improve ASR system

“ **Automatic structure discovery and parameter tuning of neural network language model based on evolution strategy ”,**

[Tanaka et al., SLT,2016]

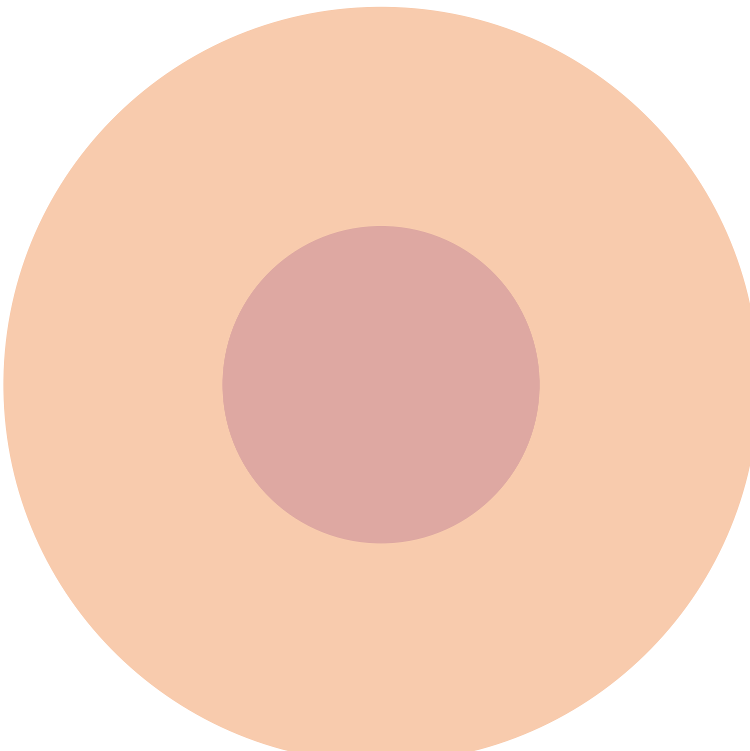
CMA-ES algorithm

- CMA-ES has shown great results in black-box optimization problems



Intuition

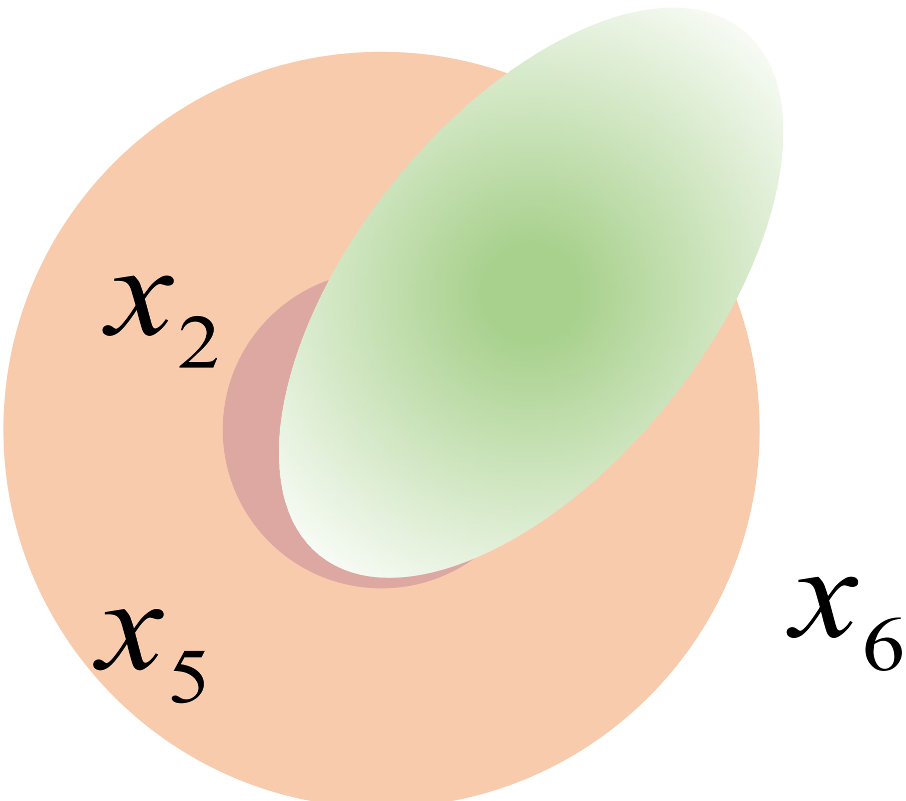
Initial (Generation 0) individual



Initialize generation 1 distribution

Intuition

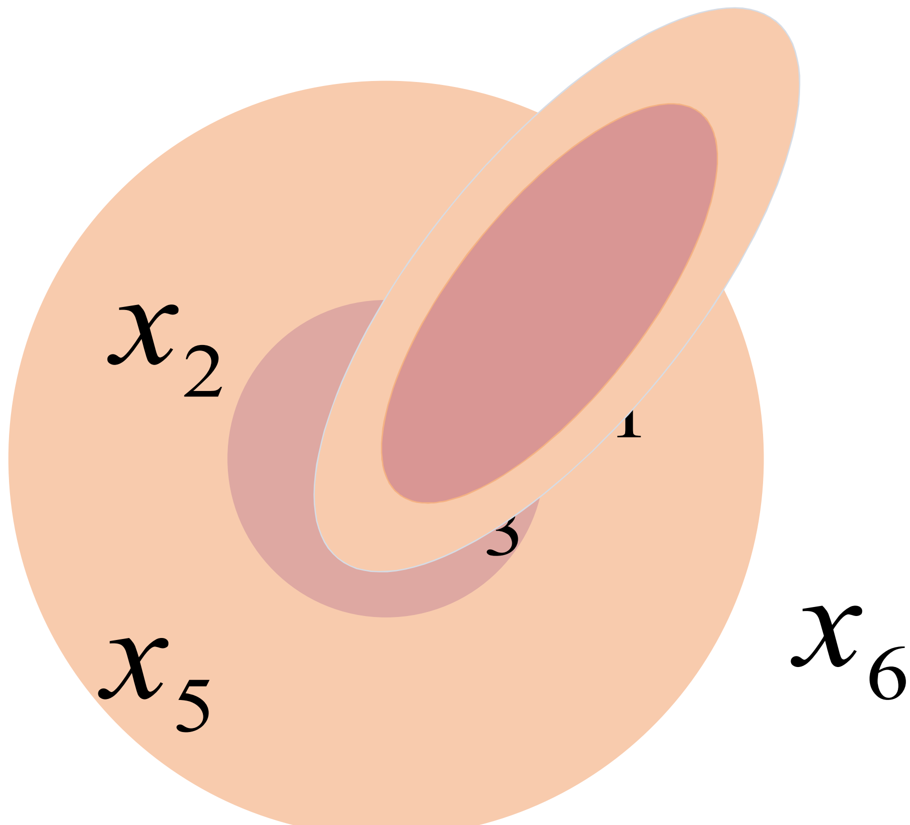
Higher performance region



Sample generation 1 individuals

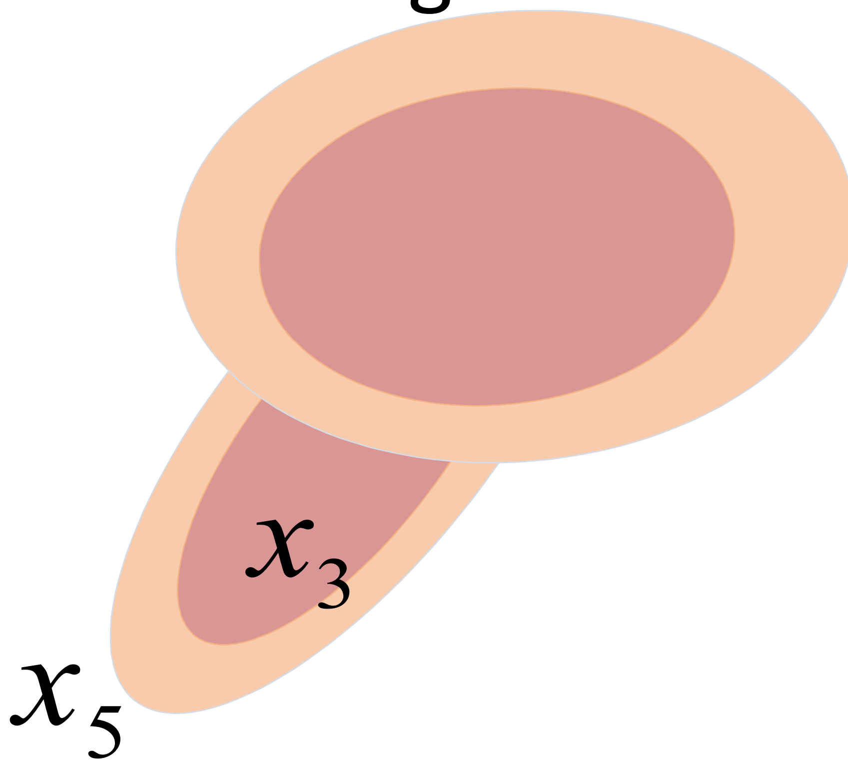
Intuition

Estimate Generation 2 distribution



Intuition

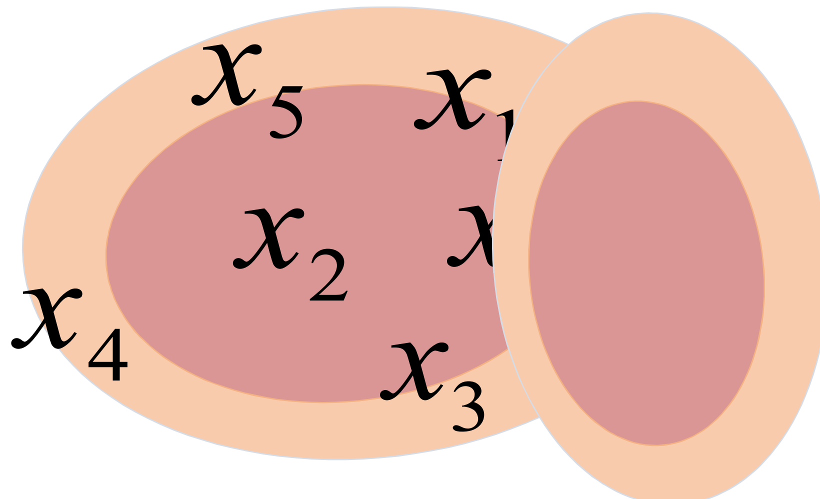
Estimate generation 3 distribution



Sample generation 2 individuals

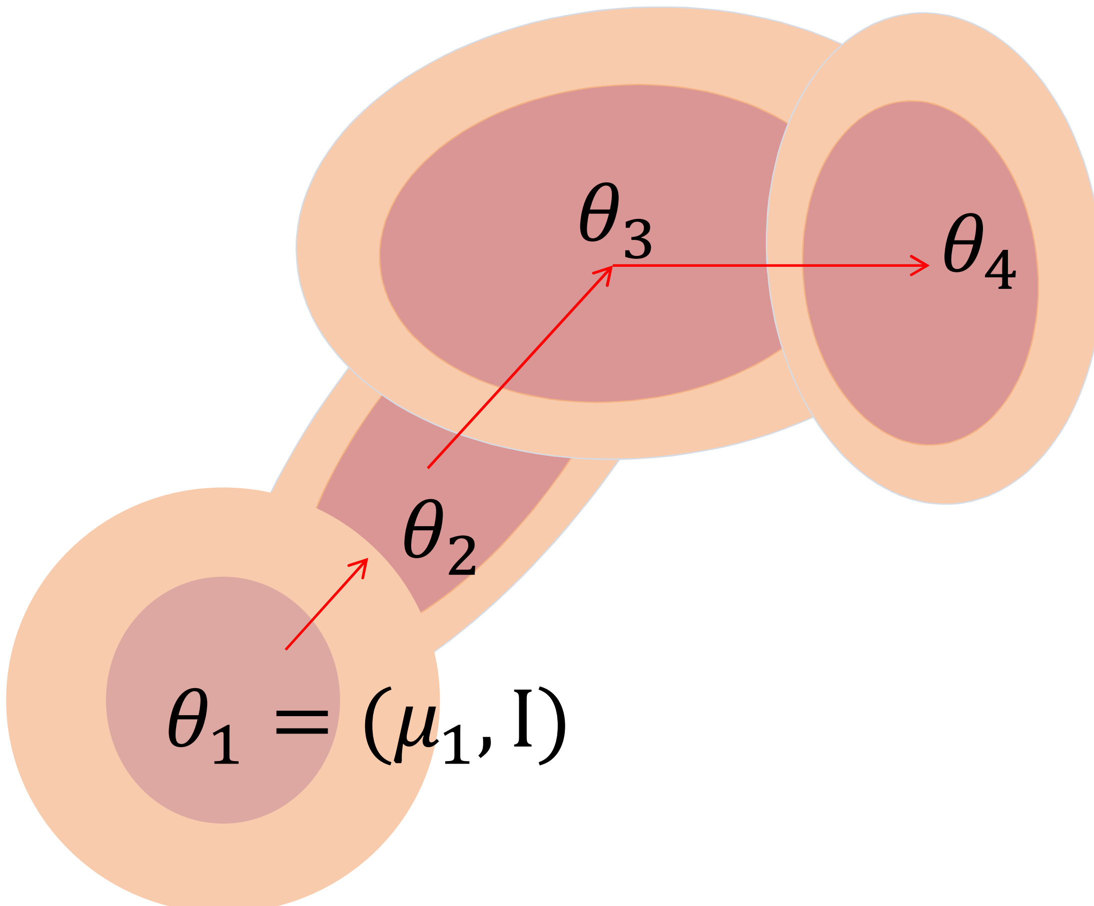
Intuition

Estimate generation 4 distribution



Sample generation 3 individuals

Intuition

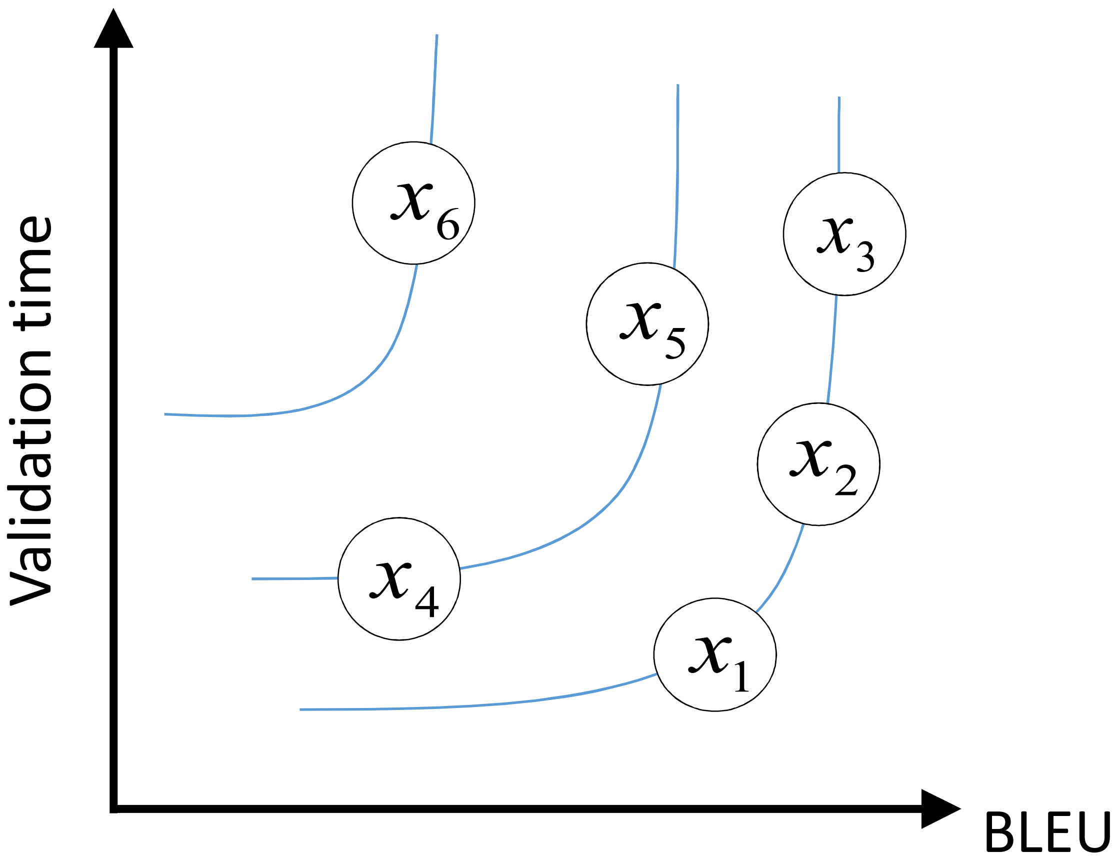


Multi-objective optimization using Pareto

- Assume that we want to maximize J objectives with respect to x jointly

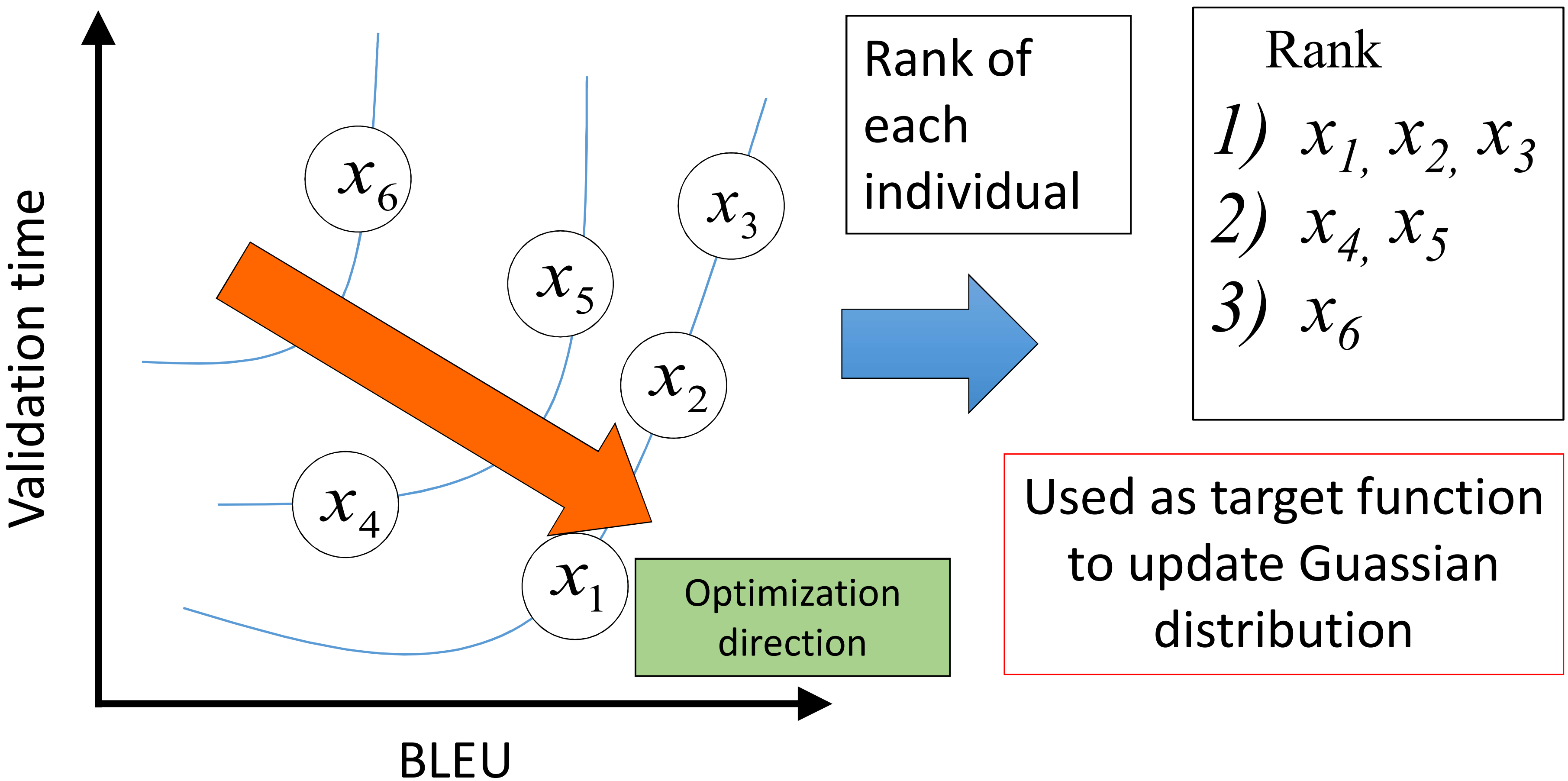
$$\mathrm{F}(x) \triangleq [f_1(x), f_2(x), \ldots, f_J(x)]$$

- As objectives might conflict with each other



BLUE and validation time opitimization

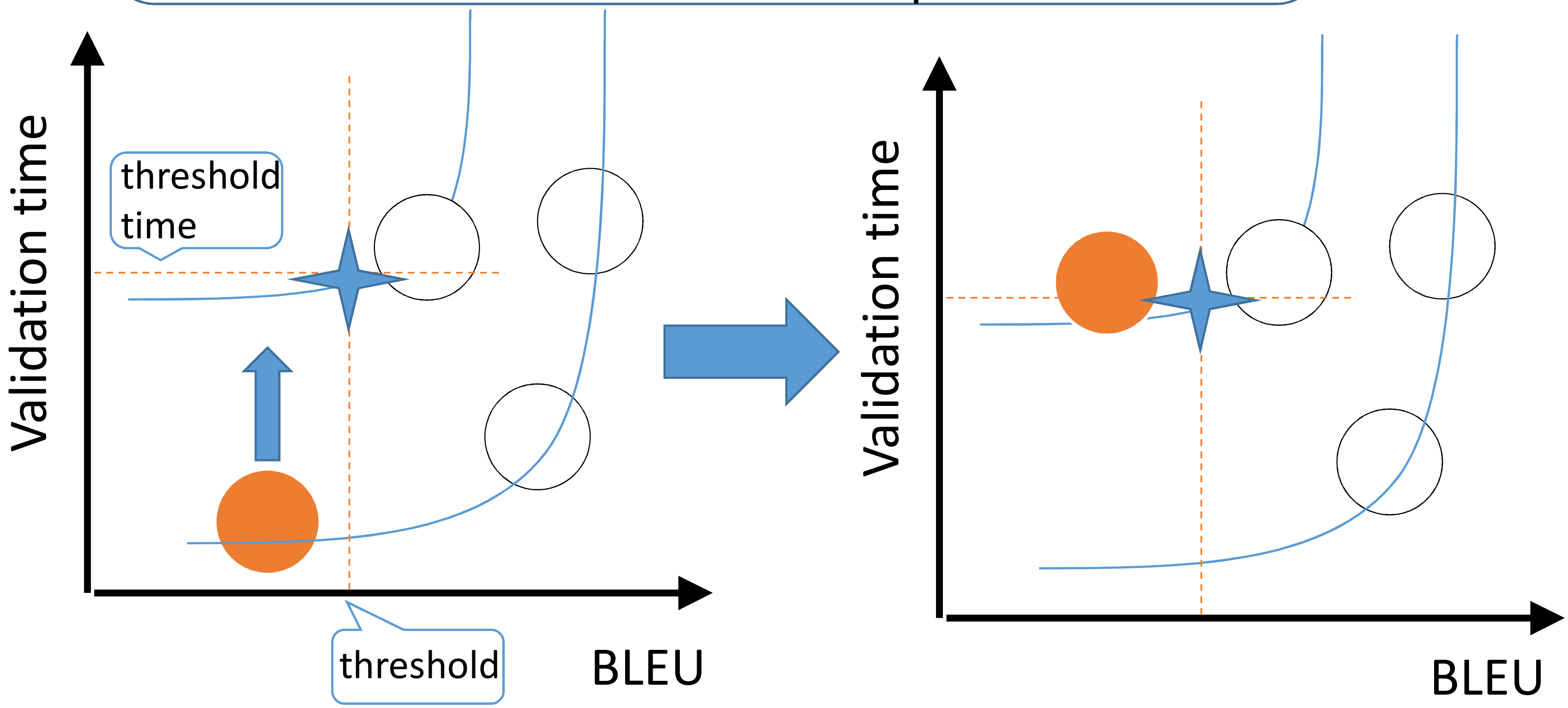
$$f_1(x) = BLUE, f_2(x) = -Validation\_time$$



Used as target function to update Guassian distribution

Practical heuristic: Threshold

- Individual with lower BLEU and smaller validation time than an initial system might have higher Pareto rank if their validation time is small
- We set a threshold to avoid this problem



Target NMT system for the tuning

- NMT system

  Nematus: attention-based neural machine translation system developed by University of Edinburgh

- Subword preprocessing

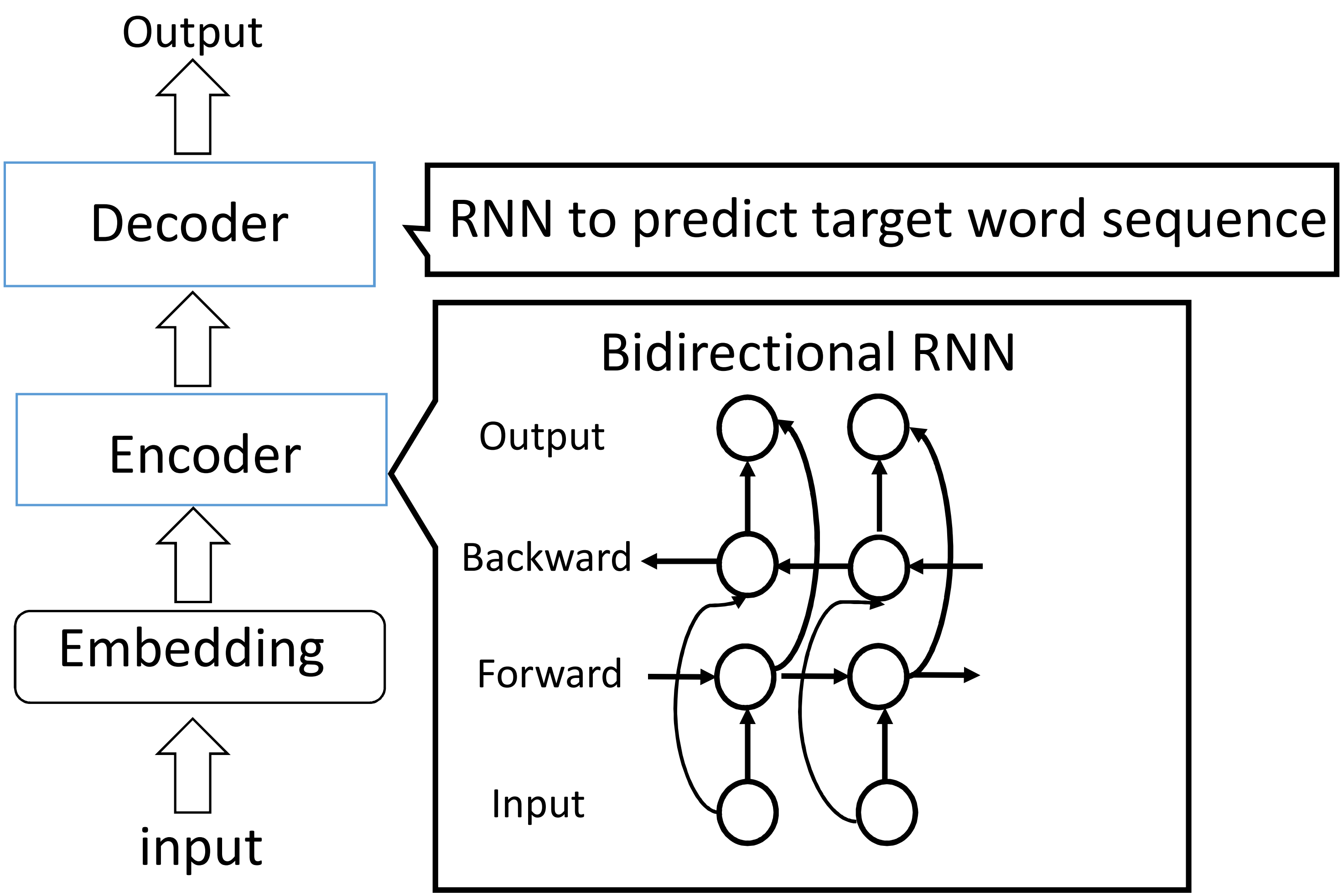
  -Words with low occurrence frequency are hard to translate

  -Using BPE (byte pair encoding) to reduce the number of distinct vocabulary items

  -The optimal # of BPE merge operation is unclear

Nematus toolkit

- Using an encoder-decoder model similar to the one proposed by Bahdanau et al.(2015), but with some implementation differences



Subword translation

- Using subword units, like morphemes or phonemes, can improve translation quality
- BPE: an algorithm to generate subword units
- # of units in BPE needs to be tuned: affects quality and time

this is the man in that house

t h i s i s t h e m a n i n t h a t h o u s e

th i s i s th e m a n i n th a t h o u s e

th i s i s th e m a n i n th a t h o u s e

Two Evolution Experiments

Single objective

- Only optimize translation accuracy

Accuracy measure : BLEU
(bilingual evaluation understudy)

- N-gram based similarity measure between translation result and reference text
- The higher the better

$$\boldsymbol{BLEU}_{(\boldsymbol{optimized})} > \boldsymbol{BLEU}_{(\boldsymbol{baseline})}$$

Multi objectives

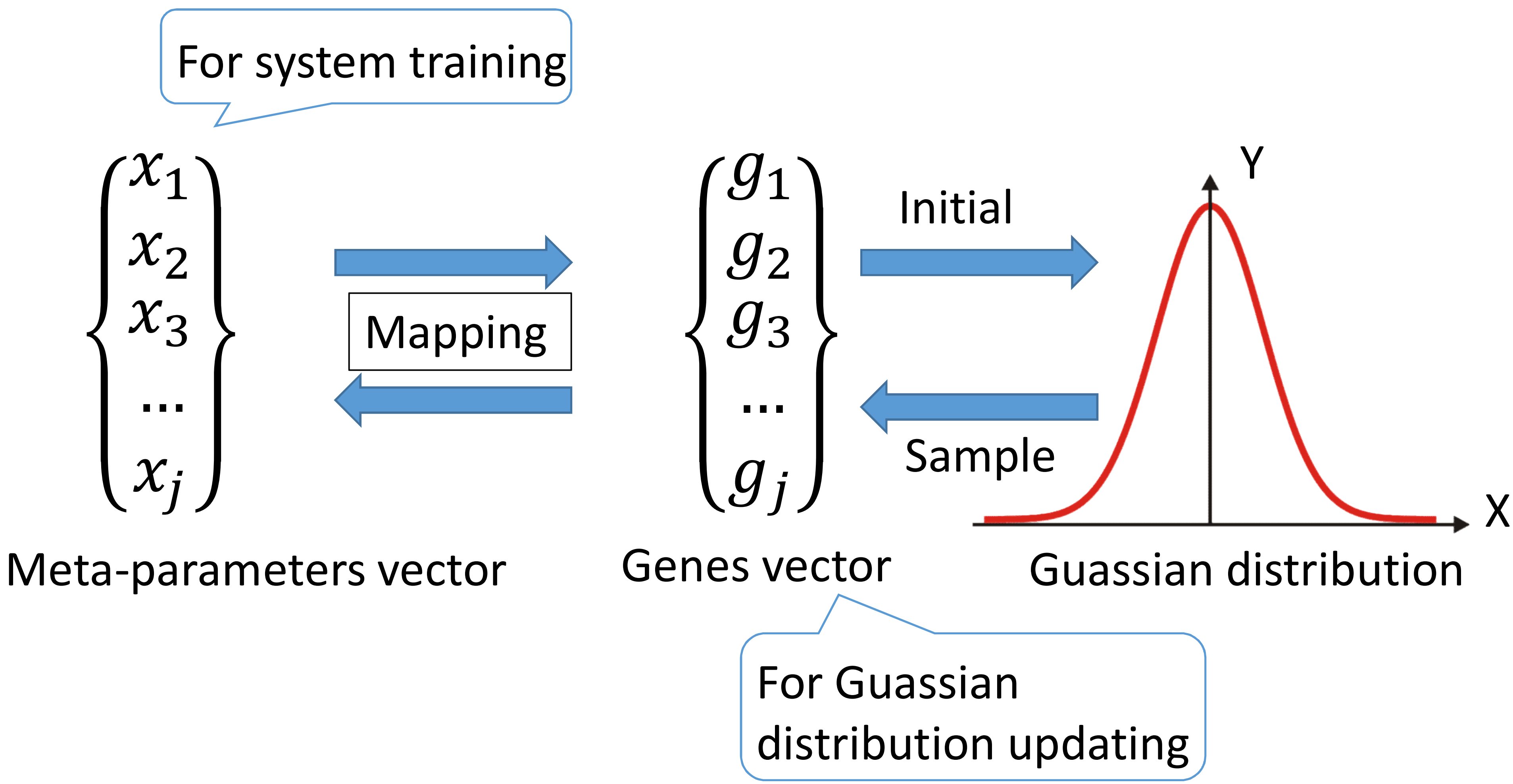
- Multi objectives means optimizing multiple objectives <u>jointly</u>, like accuracy and computational cost

Computational cost: the translation time

$$\boldsymbol{BLEU}_{(\boldsymbol{optimized})} > \boldsymbol{BLEU}_{(\boldsymbol{baseline})}$$

$$\boldsymbol{Time}_{(\boldsymbol{optimized})} < \boldsymbol{Time}_{(\boldsymbol{baseline})}$$

Gene to configuration mapping



Experimental setup

- The data comes from Kyoto free translation task (KFTT)
  -Wikipedia articles about Kyoto and Japanese culture
  -Manually translated into English by NICT
- sentences with less than 1 or more than 40 words were removed

	Articles	Sentences	Japanese words	English words
Train	14126	330k	6.09M	5.91M
Dev	15	1166	26.8k	24.3k
Test	15	1160	28.5k	26.7k

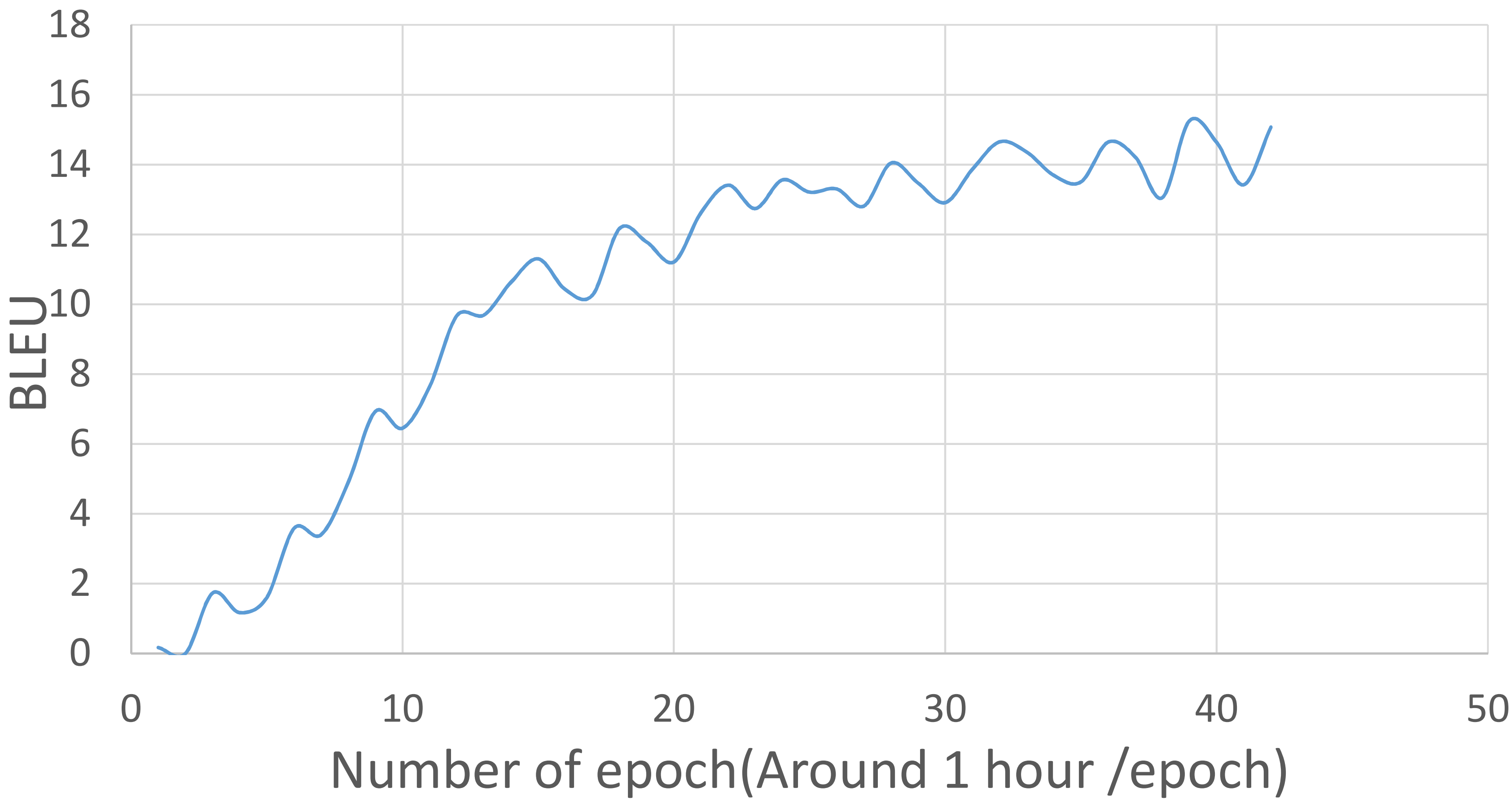
- Both sides are then broken in subword units independently using BPE

Evolution setting

training time	48hours(single), 36 hours(multi)/generation
# of generation	10 in single-obj, 5 in mul-obj
# of individuals	30
evaluation score	BLEU and validation time(seconds)
Computer machine	Tsubame 2.5 (GPU:NVIDIA K20X)
Multi-evolution threshold(BLEU)	16.5
Language	Japanese-English

Training time setting

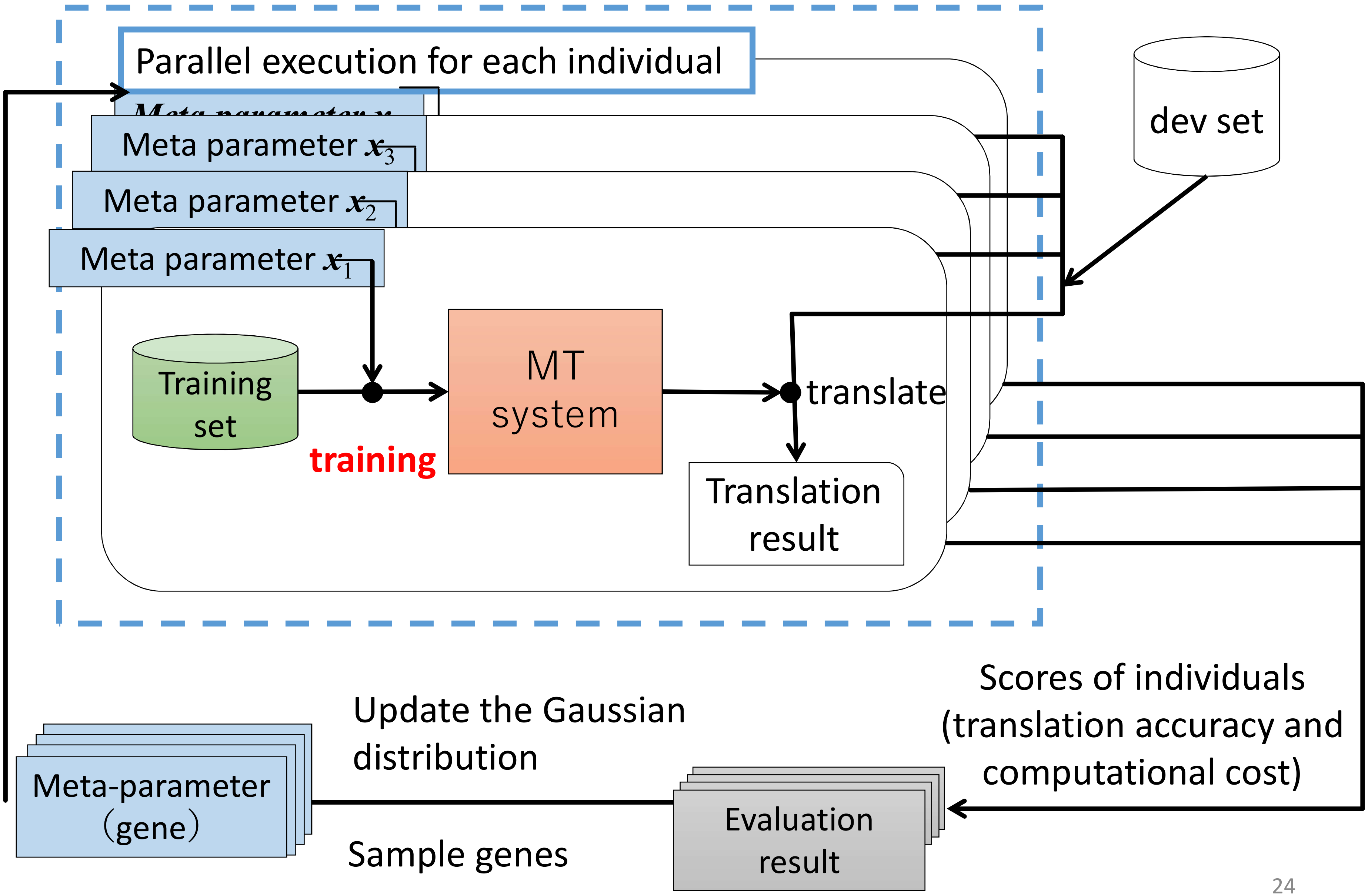
- Based on a preliminary experiment, training time is limited to 48 hours(single-objective experiment) and 36 hours(multi-objective experiment)



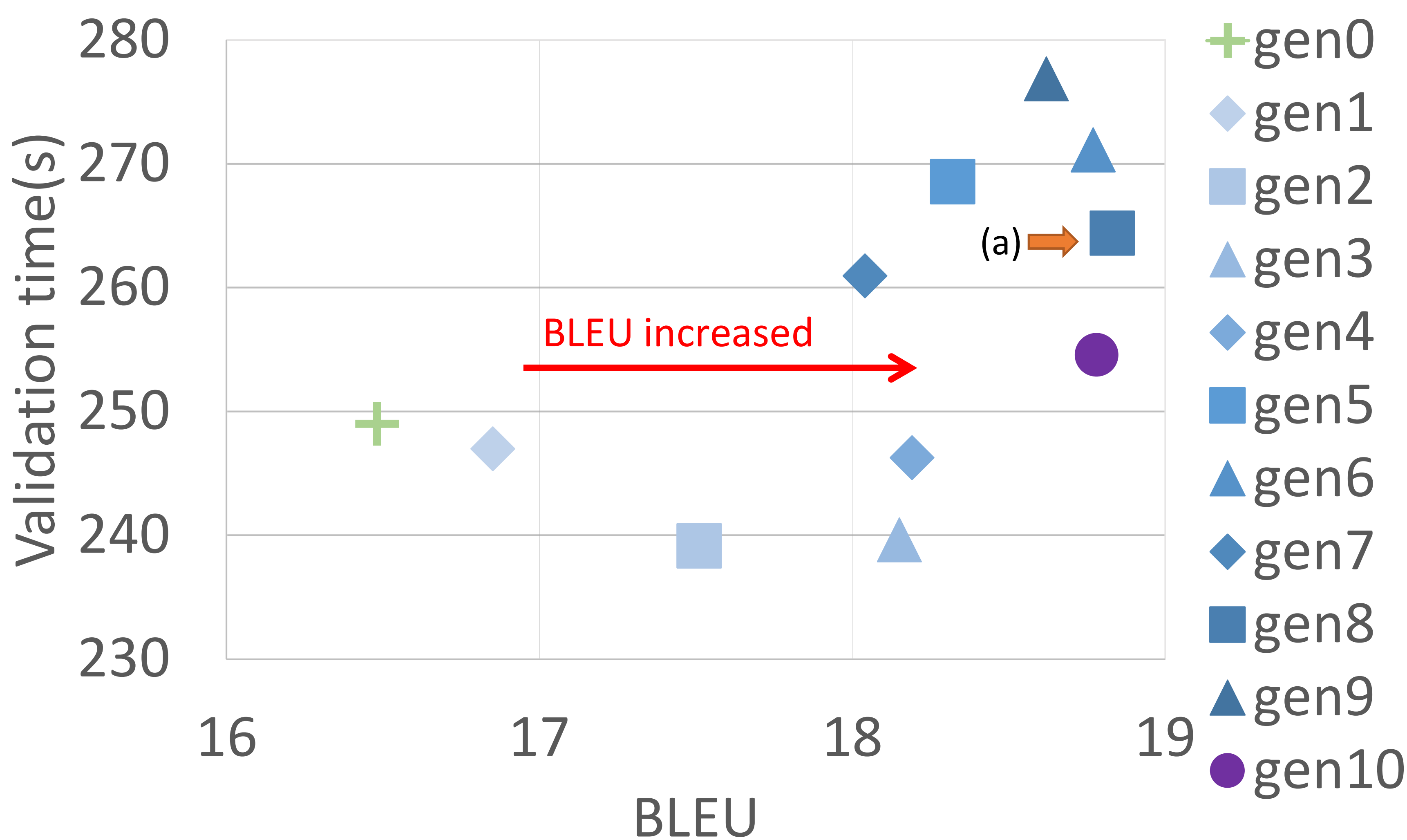
Meta-parameters to be tuned

Item	Initial value	Mapping function
BPE on source	5000	Exp
BPE on target	5000	Exp
dim of word embedding	100	Exp
dim of LSTM	400	Exp
alignment regularization	0	abs
learning rate	0.0001	abs
word embedding layer dropout	0.2	abs
hidden layer dropout	0.2	abs
source layer dropout	0.1	abs
target layer dropout	0.1	abs

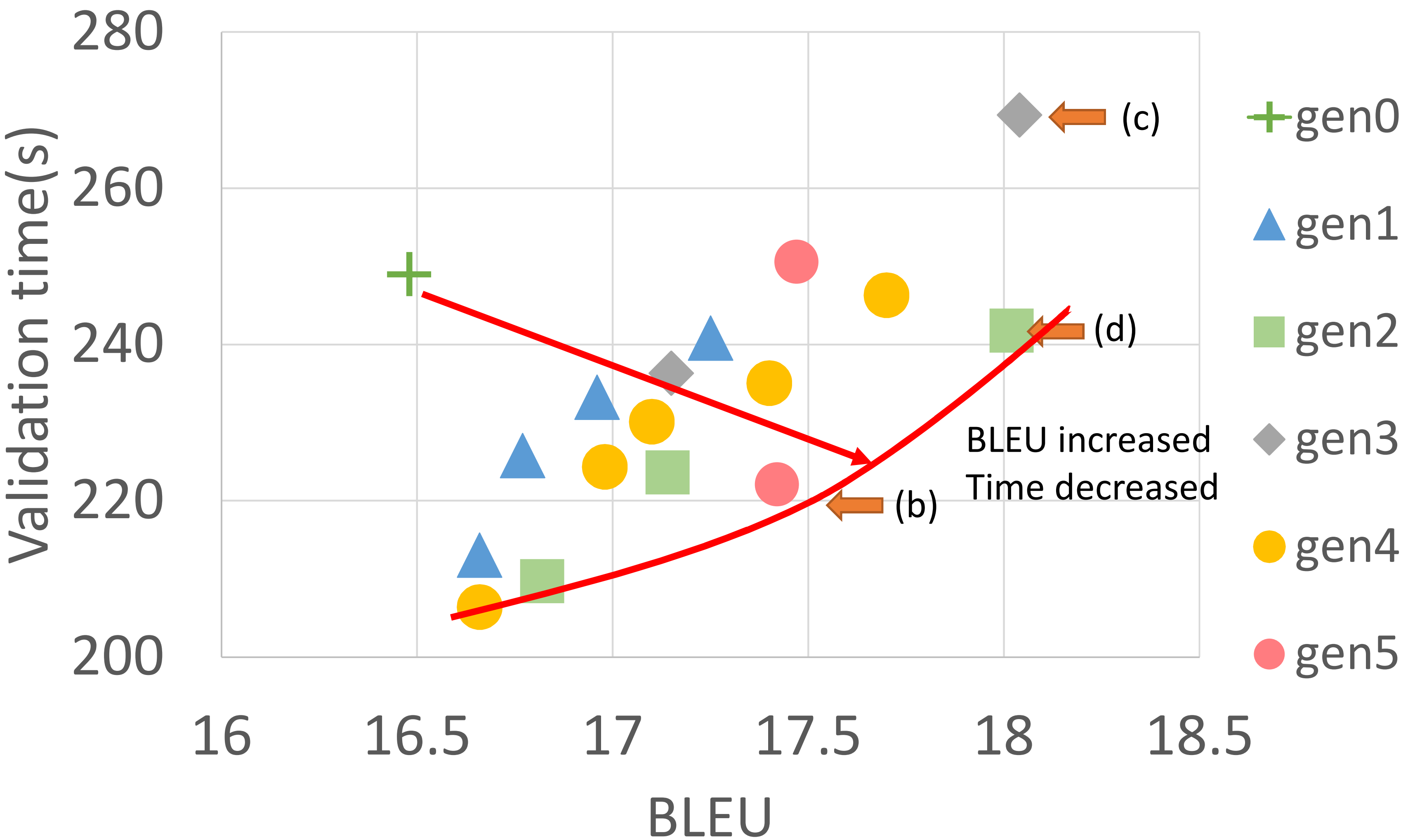
Experimental process



Single objective evolution results



Multi objectives evolution results

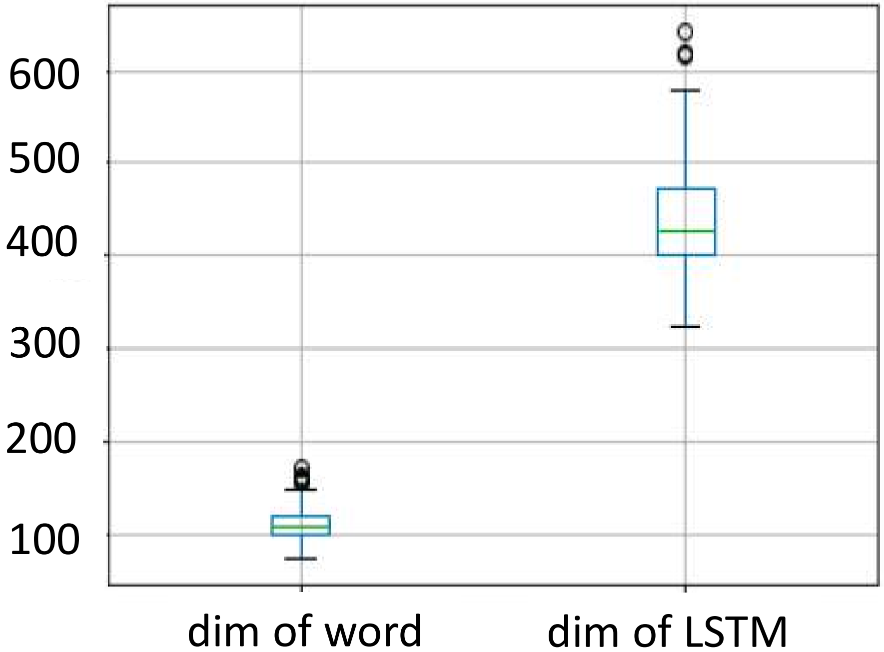


Parameter analysis

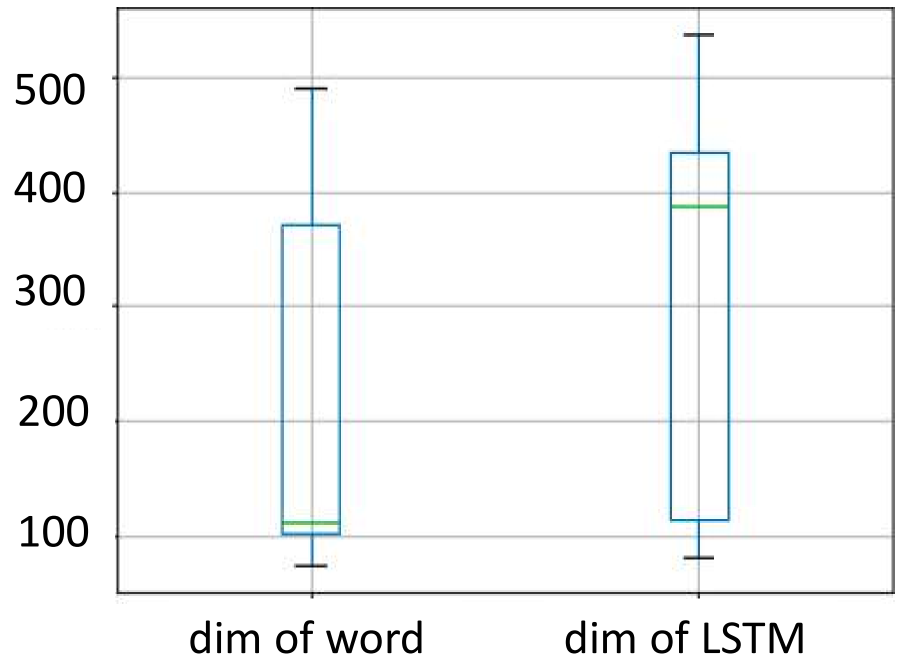
Meta-parameter	Initial value	(a)Single objective	(b)Multiple objective	(c)Multiple objective	(d)Multiple-objective
# BPE merge operation on Source(bpe_op_src)	5000	5250	5345	5011	5102
# BPE merge operation on Target(bpe_op_trg)	5000	6617	4622	5706	5877
dimension of word embedding(dim_word)	100	121	333	99	104
dimension of LSTM units(dim_lstm)	400	496	123	459	430
dev_BLEU	16.48	18.83	17.42	18.04	18.02
dev_computation time	248	264	222	269	241

Range of dimension

- The distribution of dim_lstm and dim_word in two experiments:



(a)Single objective

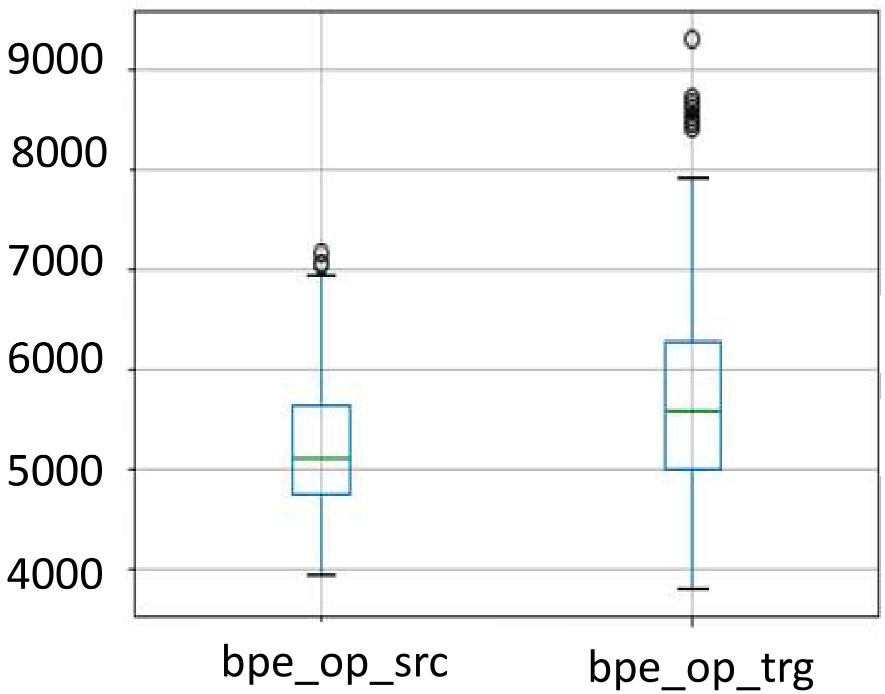


(a)Multi objective

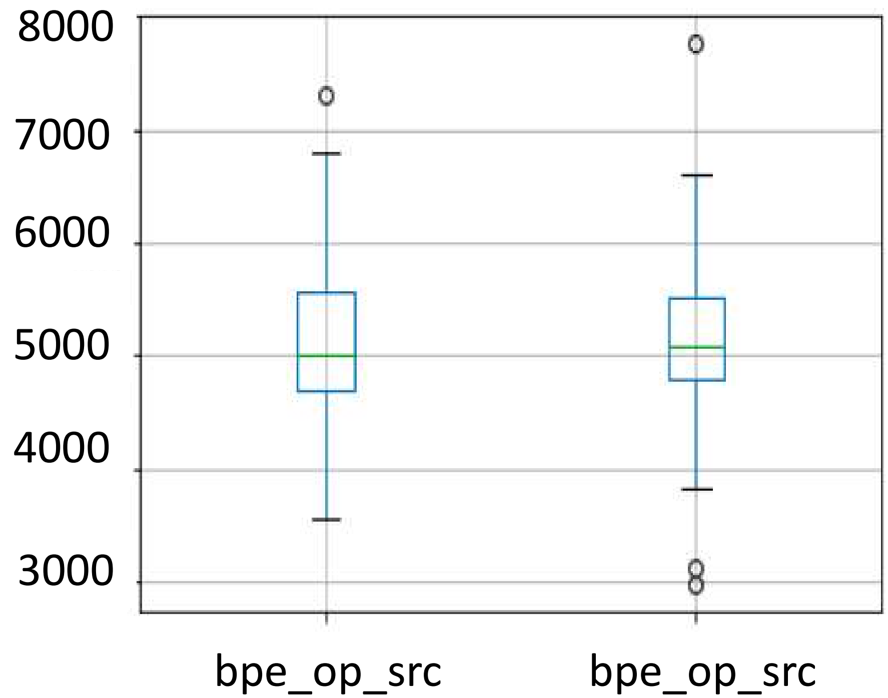
- There needs to be some more aggressive sampling in order to fully explore the meta-parameter space

Range of BPE

- The distribution of meta-parameter BPE in two experiments:



(a)Single objective,BPE



(b)Multi objective,BPE

Conclusion

- **Summary**:

  - Single-objective experiments succeeded in automatically improving the BLEU of MT system significantly
  - Multi-objective experiments needs improvement
  - Apply CMA-ES to tune NMT meta-parameter, reduce human effort

- **Next work**:

  - Adjust the setting of multi-objective experiment