Domain Adaptation for Neural Machine Translation

Kevin Duh
Johns Hopkins University
May 2019

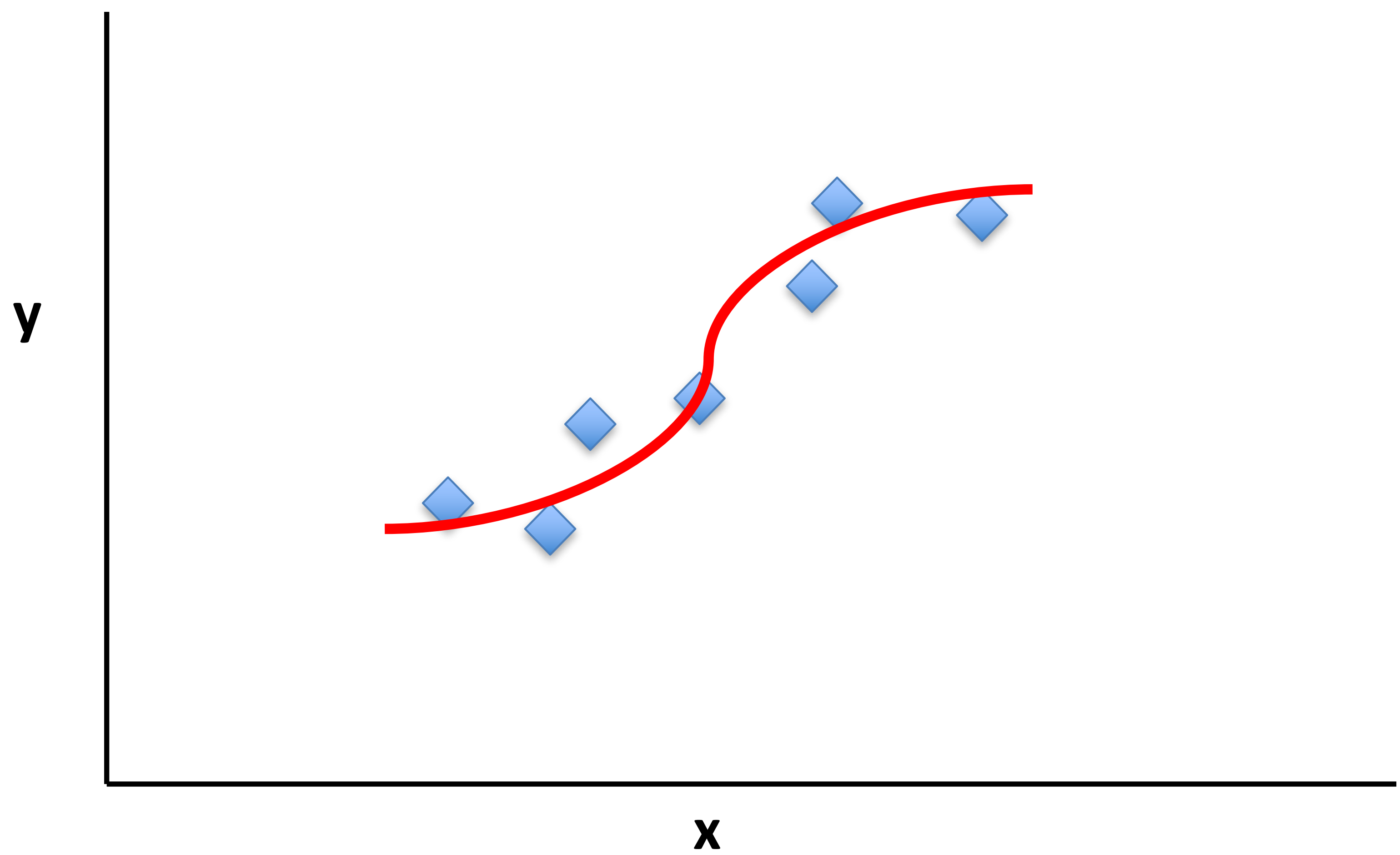
Outline

1. Problem definition
2. Survey of adaptation methods
3. Error Analysis
4. Promising Research Directions

Domain Adaptation Problem: Machine Learning Perspective

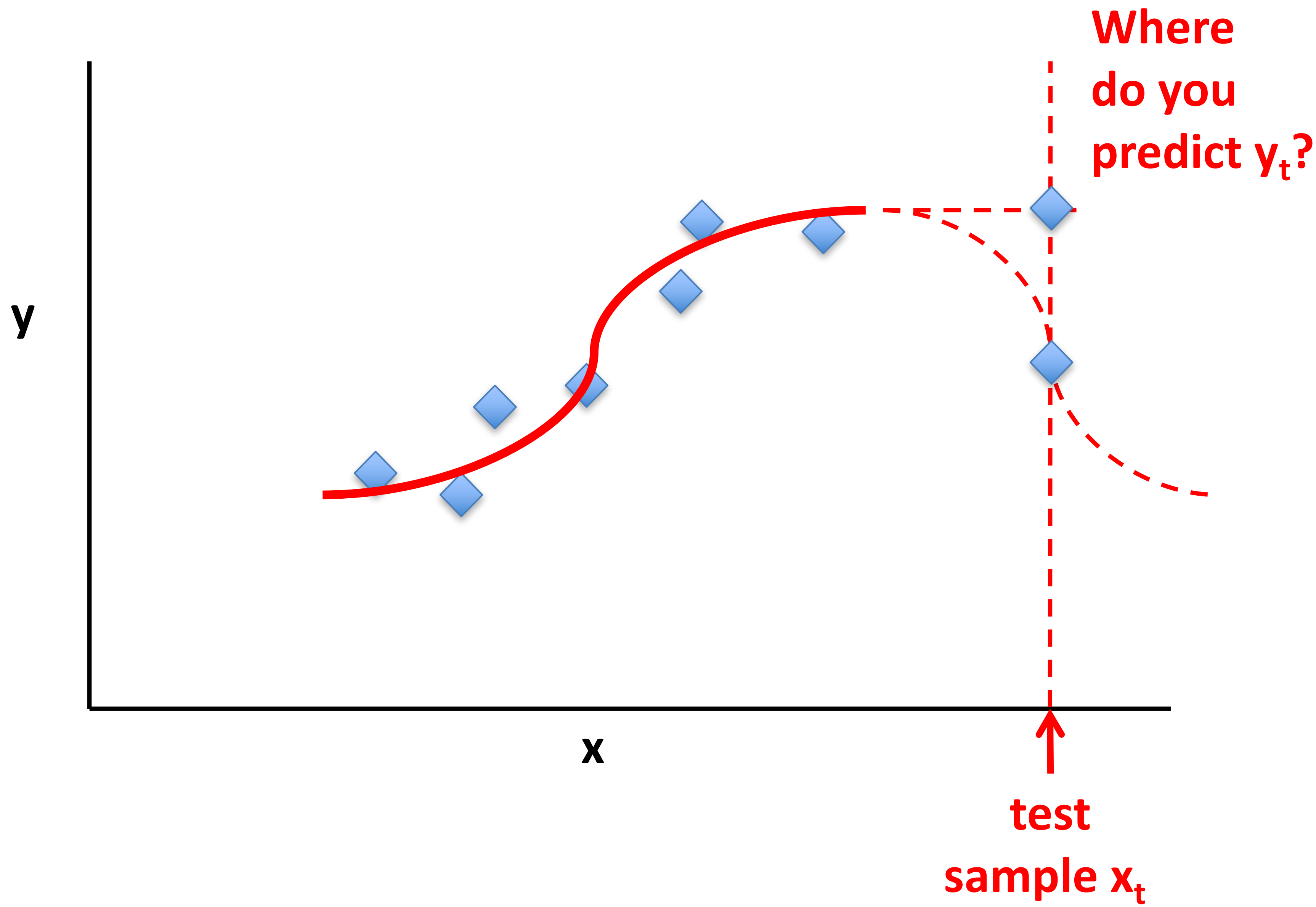
- Training data:
  - (x_1,y_1), (x_2,y_2), (x_3,y_3), …, i.i.d. samples from distribution ***D***
  - Build model $p(y|x)$
- If test data is not from ***D***, $p(y|x)$ may be operating at a space it wasn't built for.
  - Two cases for what we mean by "not from ***D***"

Visualization: Fitting p(y|x)

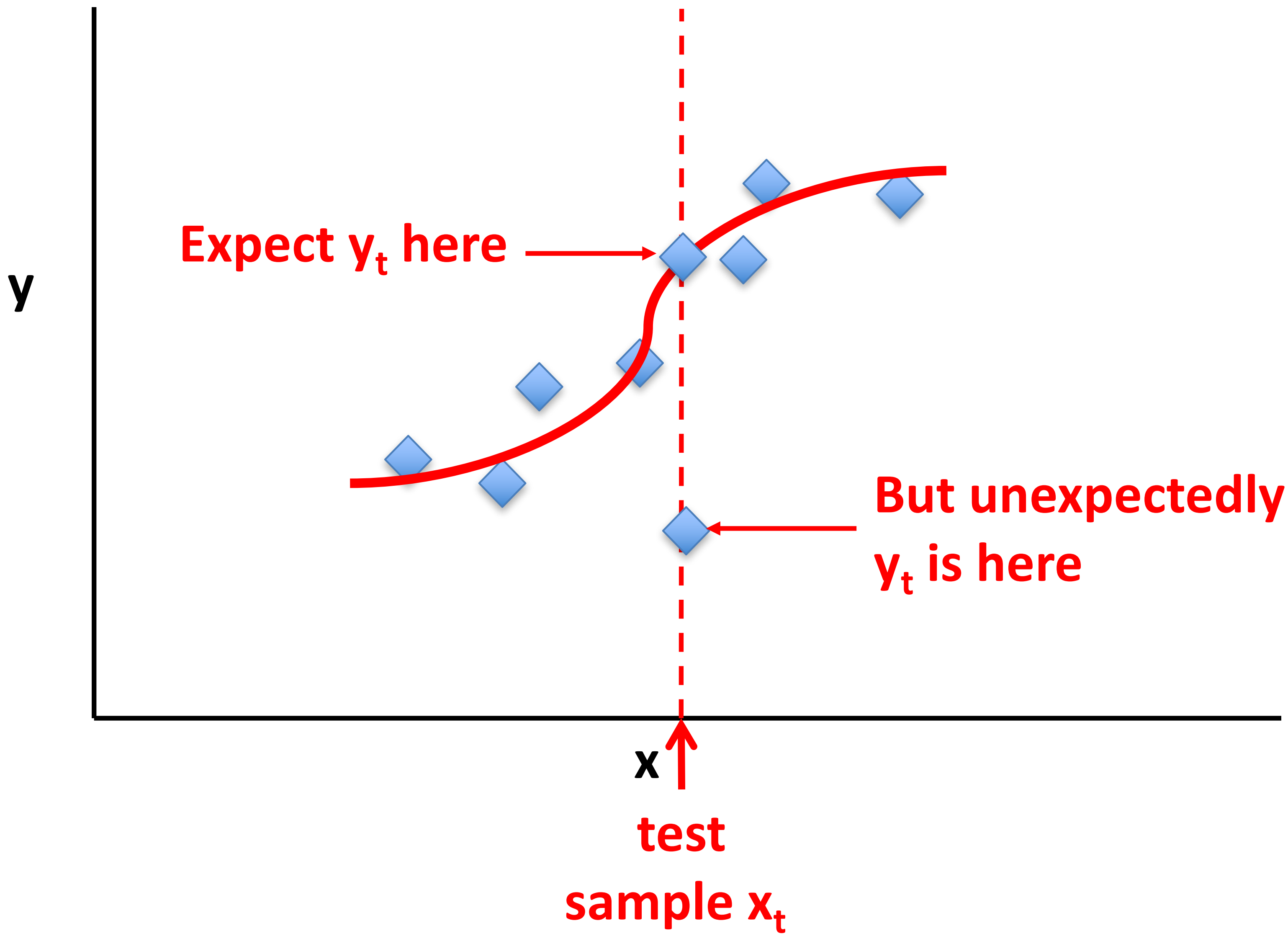


Case 1: Test is not in input domain

(Covariate Shift)



Case 2: Input-output relation changes



Examples in Machine Translation (MT)

- Domain mismatch example:
  - Training data consists of Patent sentences
  - Test sample is Social Media
- Case 1: Test is not in input domain
  - can translate technical words like “NMT”
  - no idea how to translate “OMG”
- Case 2: Input-Output relation changes
  - “CAT” translates to a word that means “Computer Aided Translation” rather than “Cute furry animal”

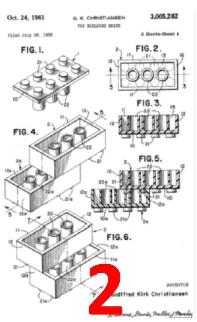
“Domain” is a fuzzy concept in practice

- Corpora differ by:
  - Topics: patents, politics, news, medicine
  - Style: formal, informal
  - Modality: written, spoken
- Often use “domain” to refer to data source:
  - e.g. Europarl, OpenSubtitles, TED, Paracrawl
- Both case 1 and case 2 mismatches occur at multiple levels: lexical, syntactic, etc.

Example sentences (case 1):

which is Patent, TED, Subtitles, Europarl?

1. We live in a digital world, but we’re fairly analog creatures.
2. The tablets exhibit improved bioavailability of the active ingredient.
3. So, um… she’s kidding.
4. Resumption of the session



2

1

Open Subtitles

3

European Parliament

4

Example bitext (case 2)

Medicine (EMEA):

if you have <u>severe depression</u>, you must not use avonex . / no debe utilizar avonex si padece una <u>depresión grave</u> .

Parliament (Europarl):

the <u>economic depression</u> in europe has lasted at least ten years . / europa sufre una <u>crisis económica</u> desde hace , al menos , diez años .

Why is Domain Adaptation an important problem in MT?

- It may be expensive to obtain training bitexts that are both large & relevant to test domain
- Often have to work with whatever we can get

Data Size

Relevance to test domain

	Small	**Large**
Irrelevant		✔
Relevant	✔	✔✔

Terminology (1 of 2)

- Example: Test domain is Social Media
- In-domain data
  - Data that is relevant to test domain: SNS corpus
- Out-of-domain data
  - Data that is less relevant to test domain: Europarl
- General-domain data
  - May use interchangeably with Out-of-Domain
  - May mean mixed corpus:
    - Europaral + Patent + TED
    - Europarl + Patent + TED + SNS

Terminology (2 of 2)

Data Size

Relevance to test domain

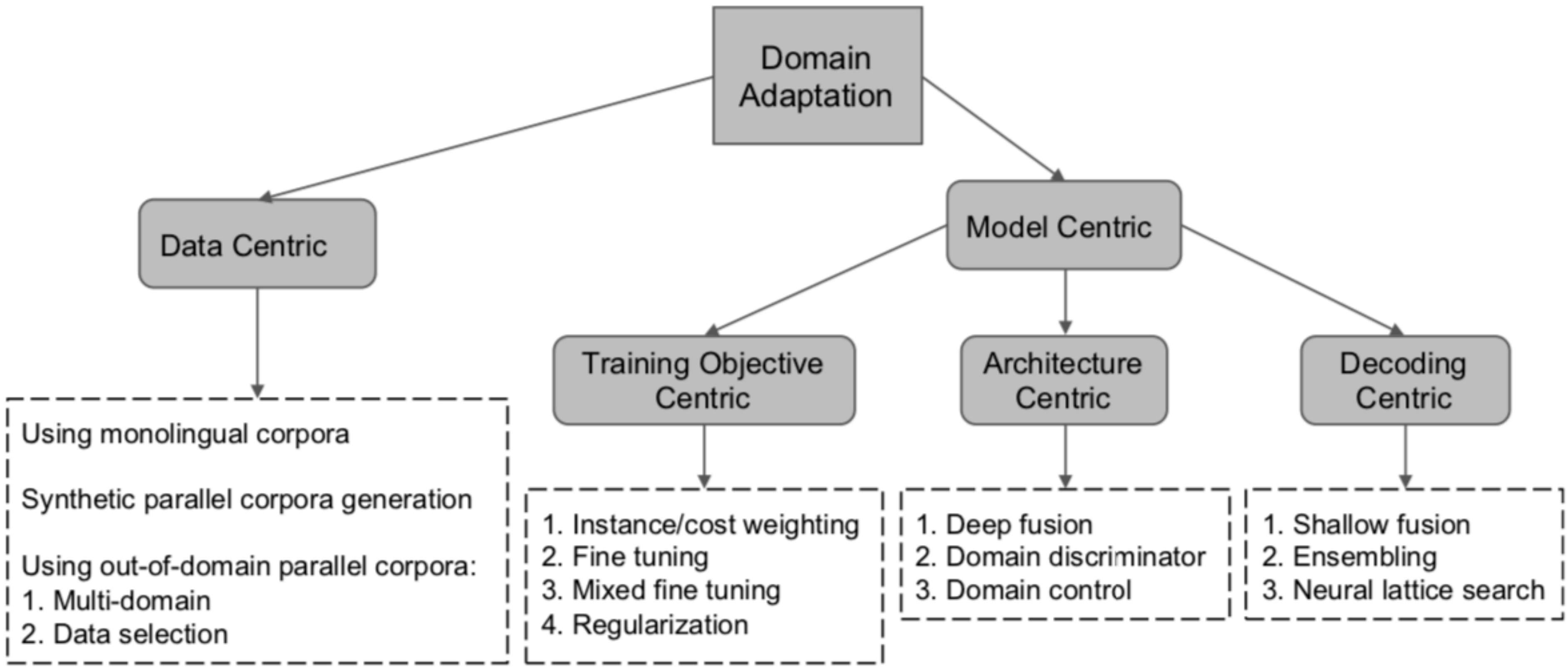
	Small	**Large**
Irrelevant		Out-of-Domain (OOD)
Relevant	In-Domain (ID)	

- Supervised adaptation methods:
  - Assumes OOD bitext & ID bitext
- Unsupervised adaptation methods:
  - Assumes OOD bitext & ID monotext

Outline

1. Problem definition
2. Survey of adaptation methods
3. Error Analysis
4. Promising Research Directions

A taxonomy of domain adaptation methods for NMT

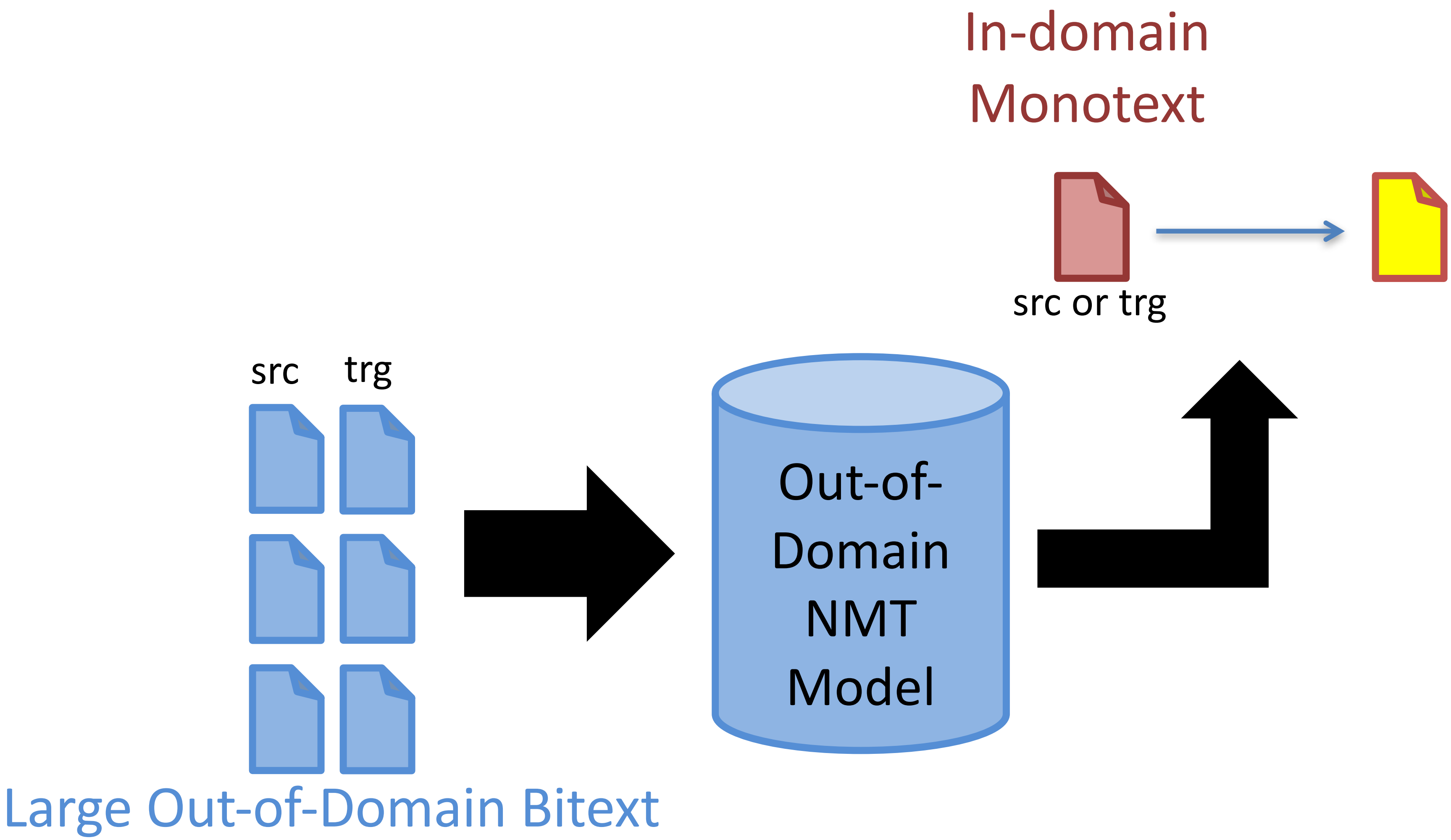


From: Chenhui Chu and Rui Wang, A Survey of Domain Adaptation for Neural Machine Translation, COLING 2018

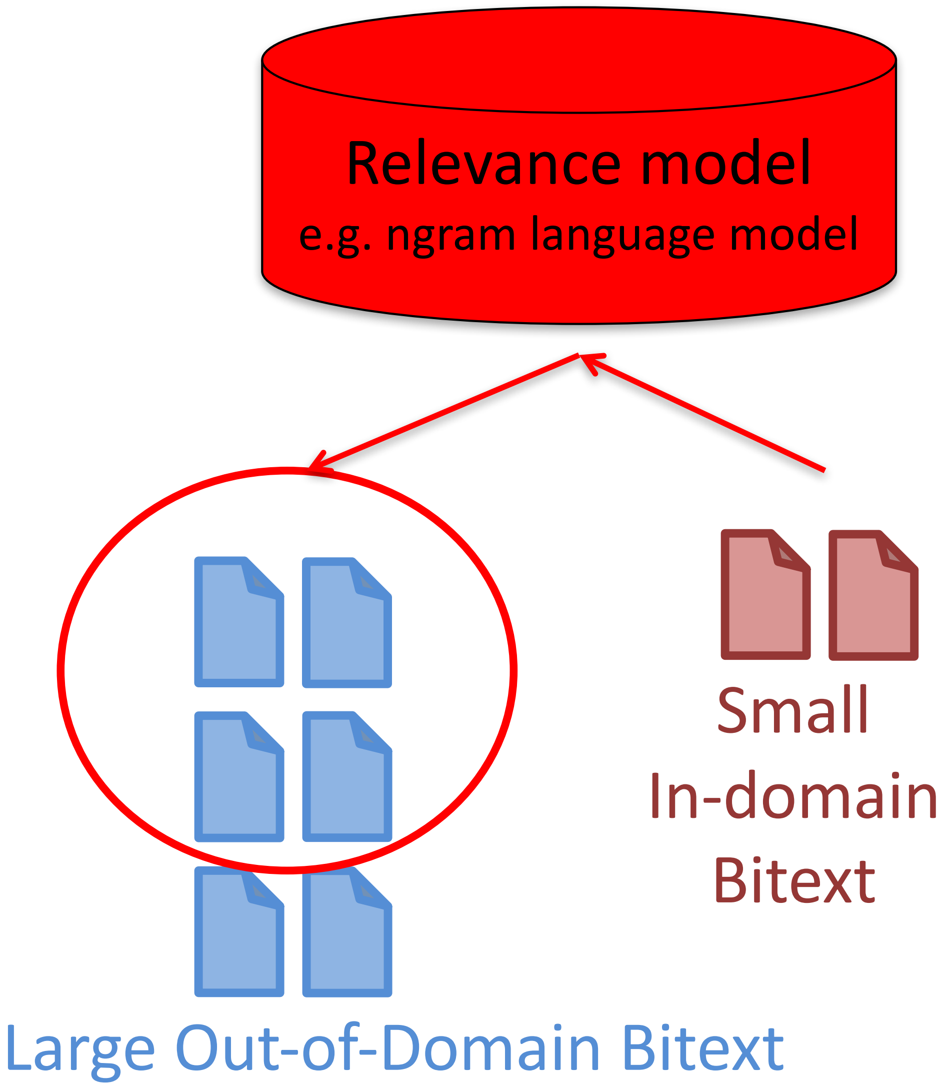
Data centric adaptation methods

Note: I'll present an assortment of methods, picked mainly to demonstrate the variety but not necessarily representative of the literature.

Synthetic Data Augmentation (forward or back translation)



Filtering Out-of-Domain Bitext for relevant data subsets (esp. for case 2)



Robert C Moore and William Lewis. Intelligent selection of language model training data. ACL 2010

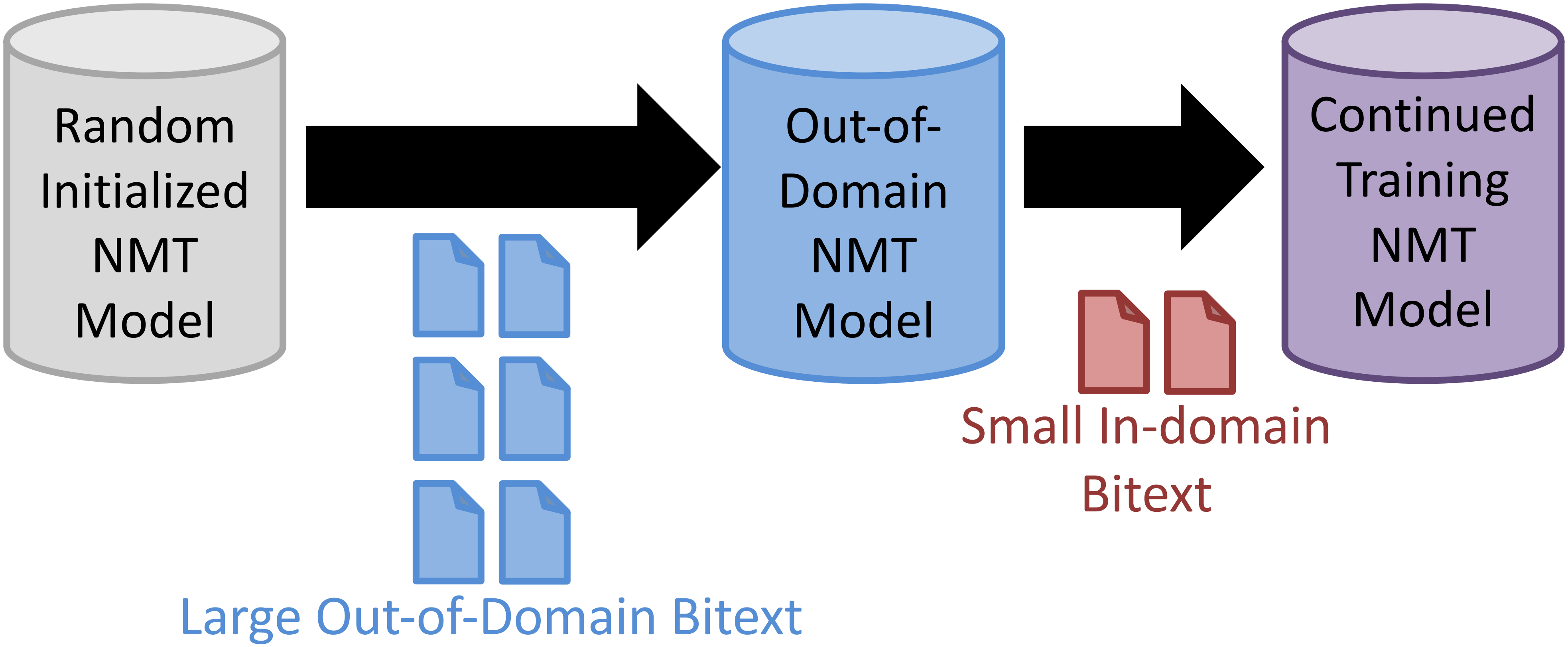
Amittai Axelrod, Xiaodong He, and Jianfeng Gao. Domain adaptation via pseudo in-domain data selection. EMNLP 2011

Kevin Duh, Graham Neubig, Katsuhito Sudoh, and Hajime Tsukada. Adaptation data selection using neural language models: Experiments in machine translation. ACL 2013

Marcin Junczys-Dowmunt. Dual conditional cross- entropy filtering of noisy parallel corpora. WMT 2018

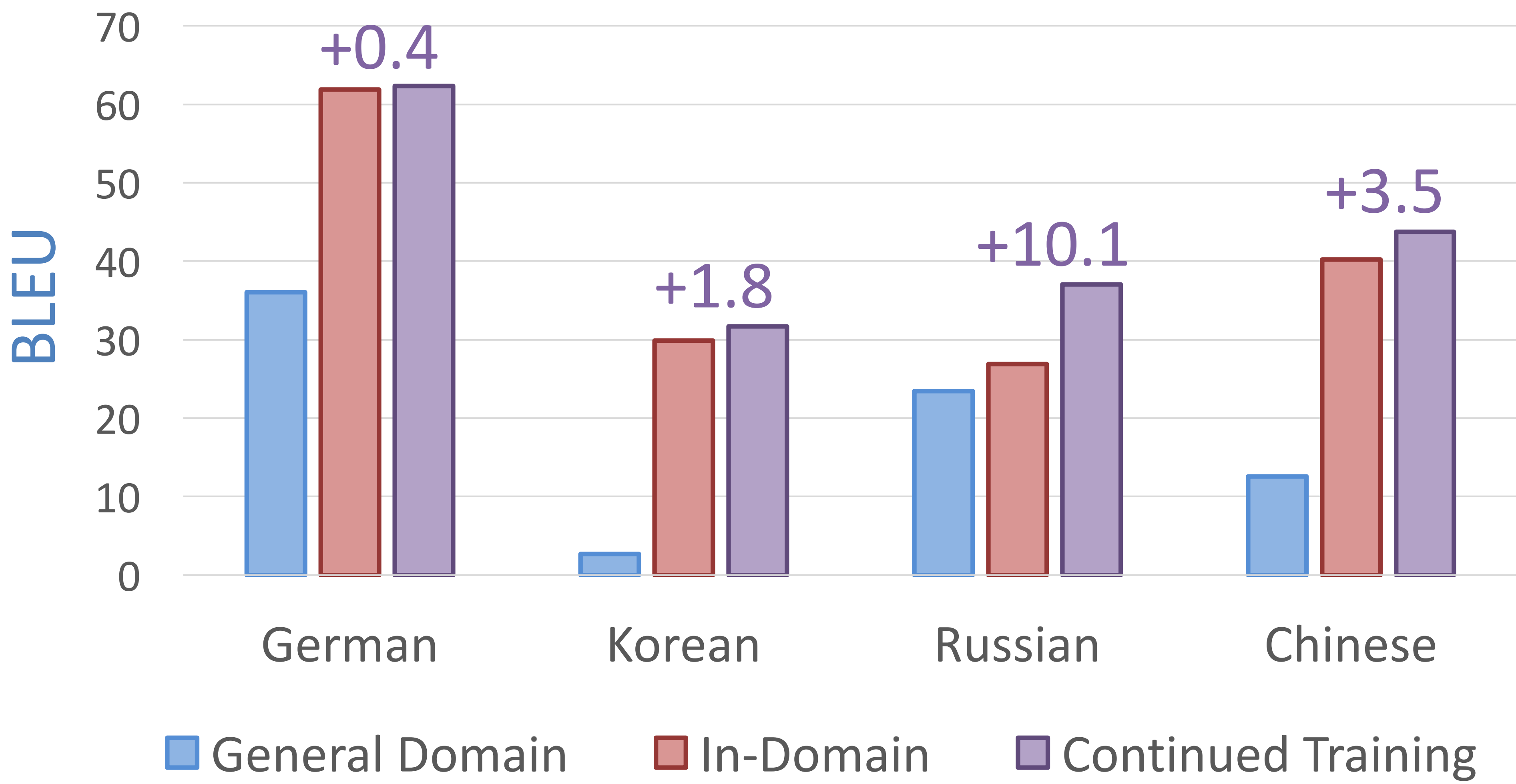
Training objective centric methods

Continued Training (a.k.a fine-tuning)



This seems to be 1st citation on NMT continued training: Minh-Thang Luong and Chris Manning. Stanford Neural Machine Translation Systems for Spoken Language Domain. IWSLT 2016

Continued Training: General Domain → Patent Domain



Results from JHU SCALE Workshop 2018: Resilient Machine Translation in New Domains

General algorithm:

1. Train model on convergence on dataset A (A=OOD bitext)
2. Continue training on dataset B (B=in-domain bitext)

Continued Training Variants:

- Details on learning rate, etc. in step 2 matters
- Adding a regularization term or fix subnetworks in step 2
  - *Antonio Valerio Miceli Barone, Barry Haddow, Ulrich Germann, and Rico Sennrich. Regularization techniques for fine-tuning in neural machine translation. EMNLP 2017*
  - *Huda Khayrallah, Brian Thompson, Kevin Duh, and Philipp Koehn. Regularized training objective for continued training for domain adaptation in neural machine translation. WNMT 2018*
  - *Brian Thompson, Huda Khayrallah, Antonios Anastasopoulos, Arya D. McCarthy, Kevin Duh, Rebecca Marvin, Paul McNamee, Jeremy Gwinnup, Tim Anderson, Philipp Koehn. Freezing Subnetworks to Analyze Domain Adaptation in Neural Machine Translation, WMT 2018*
- Different ways to mix data (e.g. A+B in step 2) or order data
  - *Chenhui Chu, Raj Dabre, and Sadao Kurohashi. An empirical comparison of domain adaptation methods for neural machine translation. ACL 2017*
  - *Marlies van der Wees, Arianna Bisazza, and Christof Monz. Dynamic data selection for neural machine translation. EMNLP 2017*
  - *Wei Wang, Taro Watanabe, Macduff Hughes, Tetsuji Nakagawa, and Ciprian Chelba. Denoising neural machine translation training with trusted data and online data selection. WMT 2018*
  - *Xuan Zhang, Pamela Shapiro, Gaurav Kumar, Paul McNamee, Marine Carpuat and Kevin Duh. Curriculum Learning for Domain Adaptation in Neural Machine Translation. NAACL 2019*
- Ensembling out-of-domain model and continued trained model:
  - *Markus Freitag and Yaser Al-Onaizan. 2016. Fast Domain Adaptation for Neural Machine Translation. ArXiV abs/1612.06897.*

Instance Weighting

Rui Wang, Masao Utiyama, Lemao Liu, Kehai Chen, Eiichiro Sumita. Instance Weighting for Neural Machine Translation Domain Adaptation. EMNLP 2017

$$J_{dw} = \lambda_{in} \sum_{(\mathbf{x},\mathbf{y}) \in \mathcal{D}_{in}} logp(\mathbf{y}|\mathbf{x}) + \sum_{(\mathbf{x}',\mathbf{y}') \in \mathcal{D}_{out}} logp(\mathbf{y}'|\mathbf{x}').$$

Boxing Chen, Colin Cherry, George Foster, Samuel Larkin. Cost Weighting for Neural Machine Translation Domain Adaptation. WNMT 2017

$$\theta^{\star} = \arg\max_{\theta} \sum_{(x,y) \in D} (1 + p_d(x)) \log p(y|x;\theta)$$

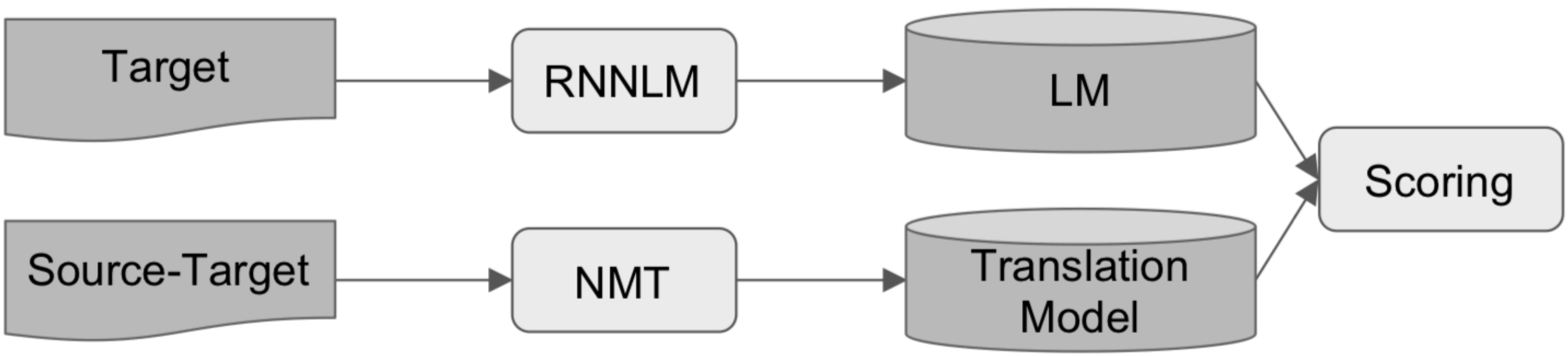
$$p_d(x) = \sigma\left(\tanh\left(W^d r_x + b^d\right)^{\top} w^d\right)$$

$$\text{where } \sigma(x) = \frac{1}{1 + \exp(-x)}$$

Architecture/Decoder Centric methods

- (I prefer to consider the two types of methods together since it is sometimes arbitrary to differentiate what's part of the whole architecture and what's only part of the decoding process)

Fusion of two models



Caglar Gulcehre, Orhan Firat, Kelvin Xu, Kyunghyun Cho, Loic Barrault, Huei-Chi Lin, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2015. On using monolingual corpora in neural machine translation. CoRR, abs/1503.03535.

Translation model vs. Language model

- Problem: Language model can be too strong

Source: Ammo muammolar hali ko'p, deydi amerikalik olim Entoni Fauchi.

Reference: But still there are many problems, says American scientist Anthony Fauci.

Baseline NMT: But there is still a lot of problems, says James Chan.

- One solution: $p(y_t \mid y_{<t}, x) = \text{softmax}(W^o \tilde{h}_t + b^o + W^\ell h_t^\ell + b^\ell)$

Averaged source word representation at decode time t

Toan Nguyen and David Chiang. Improving Lexical Choice in Neural Machine Translation. NAACL 2018 (note this paper addresses the general problem of improper lexical choice, but this is a frequent problem in domain adaptation)

Other adaptation methods

Adaptation at the token level: Subword Regularization

Subwords (_ means spaces)	Vocabulary id sequence
_Hell/o/_world	13586 137 255
_H/ello/_world	320 7363 255
_He/llo/_world	579 10115 255
_/He/l/l/o/_world	7 18085 356 356 137 255
_H/el/l/o/_/world	320 585 356 137 7 12295

Table 1: Multiple subword sequences encoding the same sentence “Hello World”

Taku Kudo, Subword Regularization: Improving Neural Network Translation Models with Multiple Subword Candidates, ACL 2018

$$\mathcal{L} = \sum_{s=1}^{|D|} \log(P(X^{(s)})) = \sum_{s=1}^{|D|} \log\Big(\sum_{\mathbf{x} \in \mathcal{S}(X^{(s)})} P(\mathbf{x}) \Big)$$

Train over different subword segmentations, randomly sampled

Domain (size)	Corpus	Language pair	Baseline (BPE)	Proposed (SR)
Web	IWSLT15	en → vi	13.86	17.36*
(5k)		vi → en	7.83	11.69*
		en → zh	9.71	13.85*
		zh → en	5.93	8.13*
	IWSLT17	en → fr	16.09	20.04*
		fr → en	14.77	19.99*
	WMT14	en → de	22.71	26.02*
		de → en	26.42	29.63*

Outline

1. Problem definition
2. Survey of adaptation methods
3. Error Analysis
4. Promising Research Directions

The truth is, it’s hard to analyze MT errors

- It was hard to pinpoint why translation was incorrect in the SMT days
- It’s perhaps even harder for NMT
- But we try anyway. At least it gives a way to think about the problem.

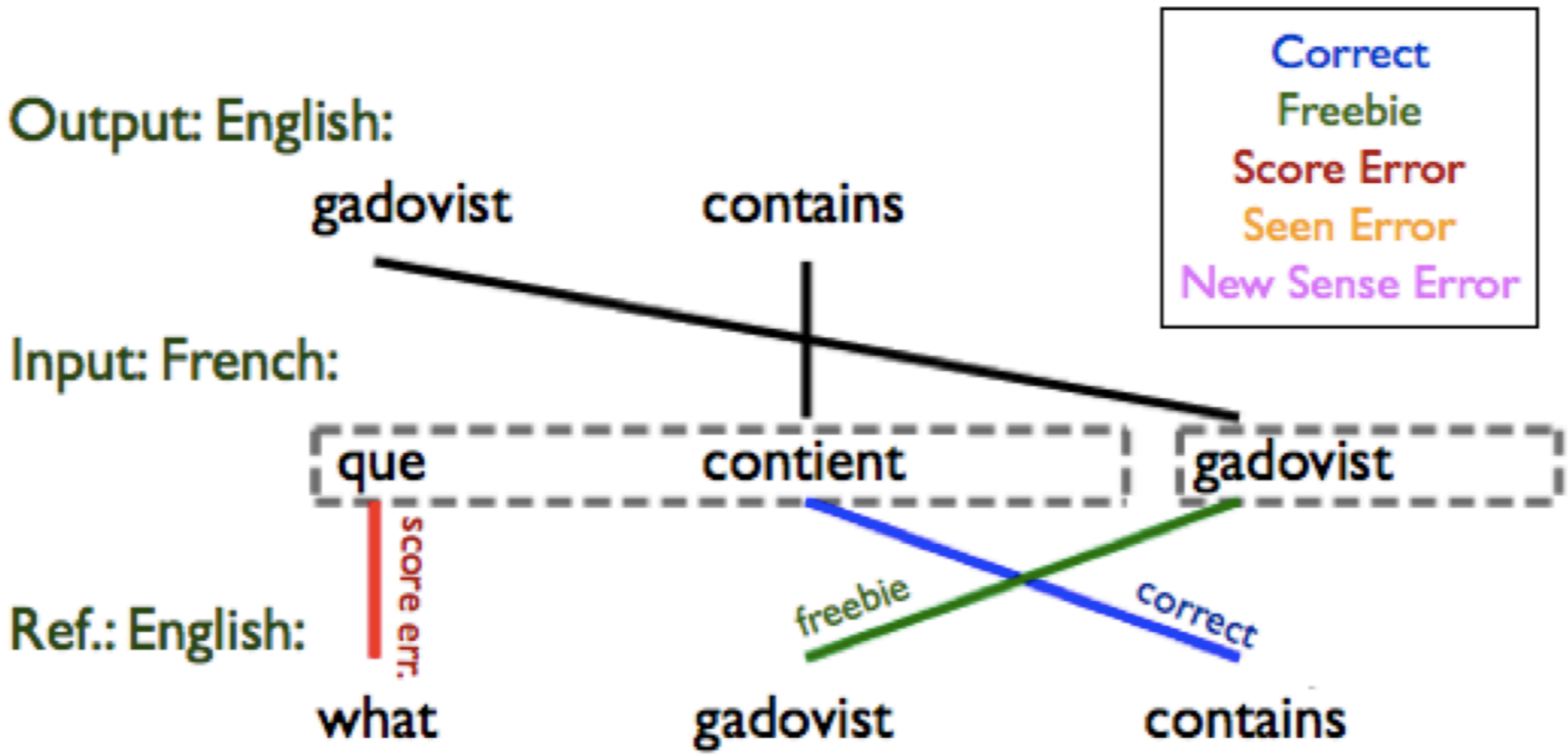
S4 Analysis

(originally developed for SMT)

- **SEEN error**: Never seen this source word before in the training data (case 1 in 1st part of this talk)
- **SENSE error**: The source word appears in the training data, but is not used in this sense. (e.g. case 2)
- **SCORE error**: The source word and its translation appears in the training data, but the correct translation is scored lower
- **SEARCH error**: The correct translation is scored higher, but somehow got lost in the search process

Ann Irvine, John Morgan, Marine Carpuat, Hal Daume´ III, and Dragos Munteanu. 2013. Measuring machine translation errors in new domains. Transactions of the Association for Computational Linguistics (TACL)

S4 Analysis
(requires reference and alignment)



Ann Irvine, John Morgan, Marine Carpuat, Hal Daume´ III, and Dragos Munteanu. 2013. Measuring machine translation errors in new domains. Transactions of the Association for Computational Linguistics (TACL)

S4 Analysis
(for NMT?)

- First, run external word aligner to determine correct/incorrect words (but how much can we trust this?)

- **SEEN error**:
  - Check out-of-vocabulary words on source side
- **SENSE error**:
  - For a given source word **f**, check if the desired translation never appears in the target side of training bitext where **f** appears?
- **SCORE error**:
  - When none of the above is true?
- **SEARCH error**:
  - Not sure how to check besides increasing beam, but..

Fluently Inadequate Translations

- Source: (TEDtalk) 乌鸦 父母 还 教会 自己 的 孩子 这样 的 技巧 呢 。
- Un-adapted system output: I'm afraid I'm not going to have to go to bed.
- Gloss: Crow parents seem to be teaching their young these skills.
- Adapted system output: And their parents also taught their children how to do it.
- Translations that are fluent but have nothing to do with the source are **very dangerous**!

Marianna J. Martindale, Marine Carpuat, Kevin Duh and Paul McNamee. Identifying Fluently Inadequate Output in Neural and Statistical Machine Translation. MT Summit 2019

Outline

1. Problem definition
2. Survey of adaptation methods
3. Error Analysis
4. Promising Research Directions (my opinion)

Personalized Adaptation, e.g. for Computer Assisted Translation (CAT)

- CAT presents many interesting opportunities for research (with real user impact!)
- Example interface at Lilt.com:

Paul Michel, Graham Neubig. Extreme Adaptation for Personalized Neural Machine Translation. ACL 2018
Sachith Sri Ram Kothur, Rebecca Knowles and Philipp Koehn. Document-Level Adaptation for Neural Machine Translation. WNMT 2018

Adaptation to New Languages

- Given bitext in language pairs A->B, C->D
  - Build a translator for A->D
  - Build a translator for E->B where E is related language to A
  - Assumes some shared representation, can use continued training, etc.
- Crazy idea but potentially large impact

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. Transfer learning for low-resource neural machine translation. EMNLP 2016; Graham Neubig and Junjie Hu. Rapid Adaptation of Neural Machine Translation to New Languages. EMNLP 2018

Understanding adaptation errors as a way to understand NMT behavior

- Domain Adaptation provides a good testbed for understanding overfitting, etc.
- What triggers a fluently inadequate translation?
- Why does catastrophic forgetting happen?

Thompson, et. al. NAACL 2019; Saunders, et. al. ACL 2019; Kirpatrick, et. al. PNAS 2017

Better adaptation algorithms

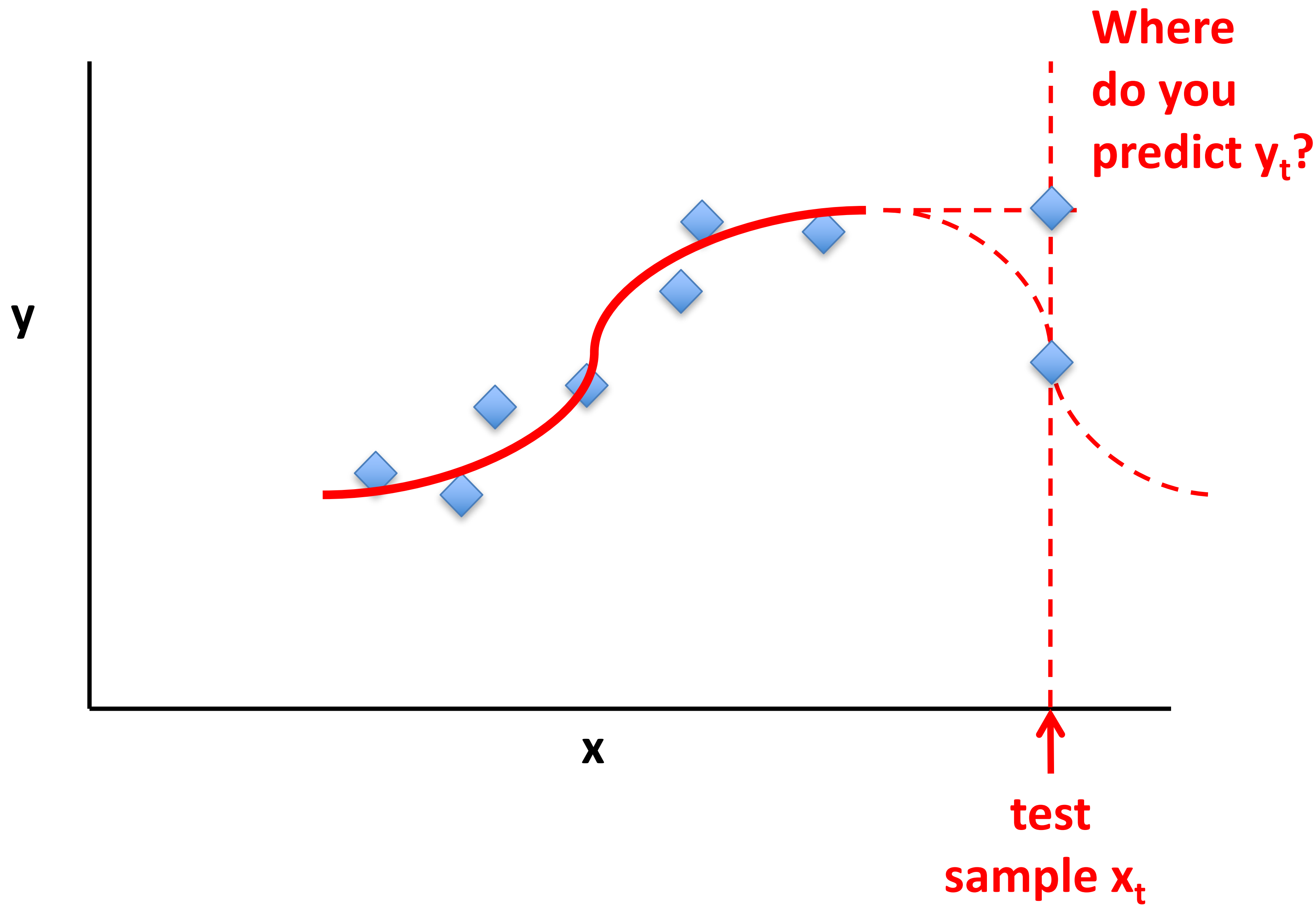
- Many opportunities for new ideas in NMT, e.g.
- Batch vs. Online setup
  - Online setup makes curriculum learning easier
- Easy to design new architectures
  - Multitask learning for sharing parameters & data
- What ideas to borrow from SMT?
  - Modularization of lexical choice, reordering, syntax, etc.

Re-cap

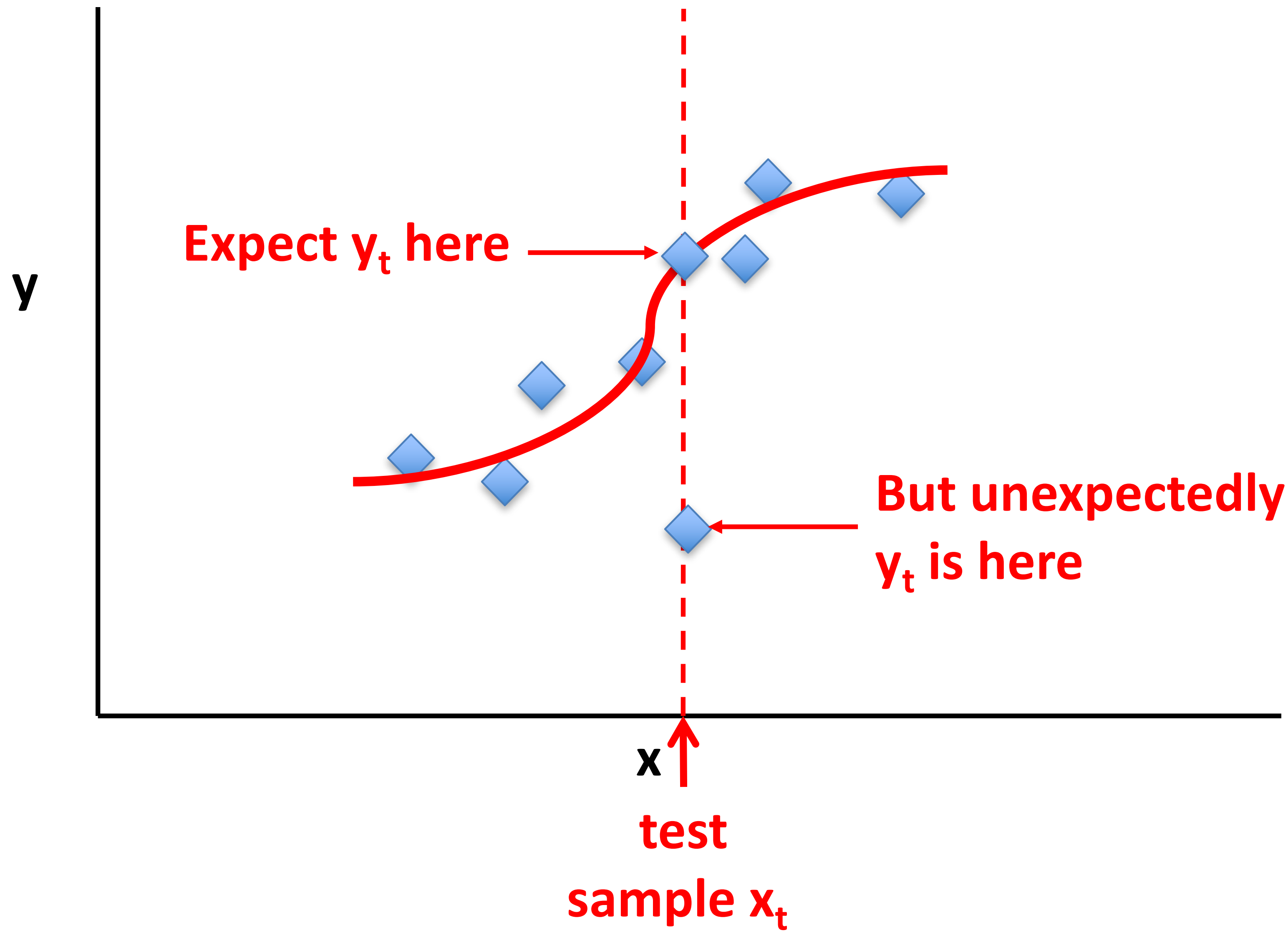
1. Problem definition
2. Survey of adaptation methods
3. Error Analysis
4. Promising Research Directions

Case 1: Test is not in input domain

(Covariate Shift)



Case 2: Input-output relation changes



Why is Domain Adaptation an important problem in MT?

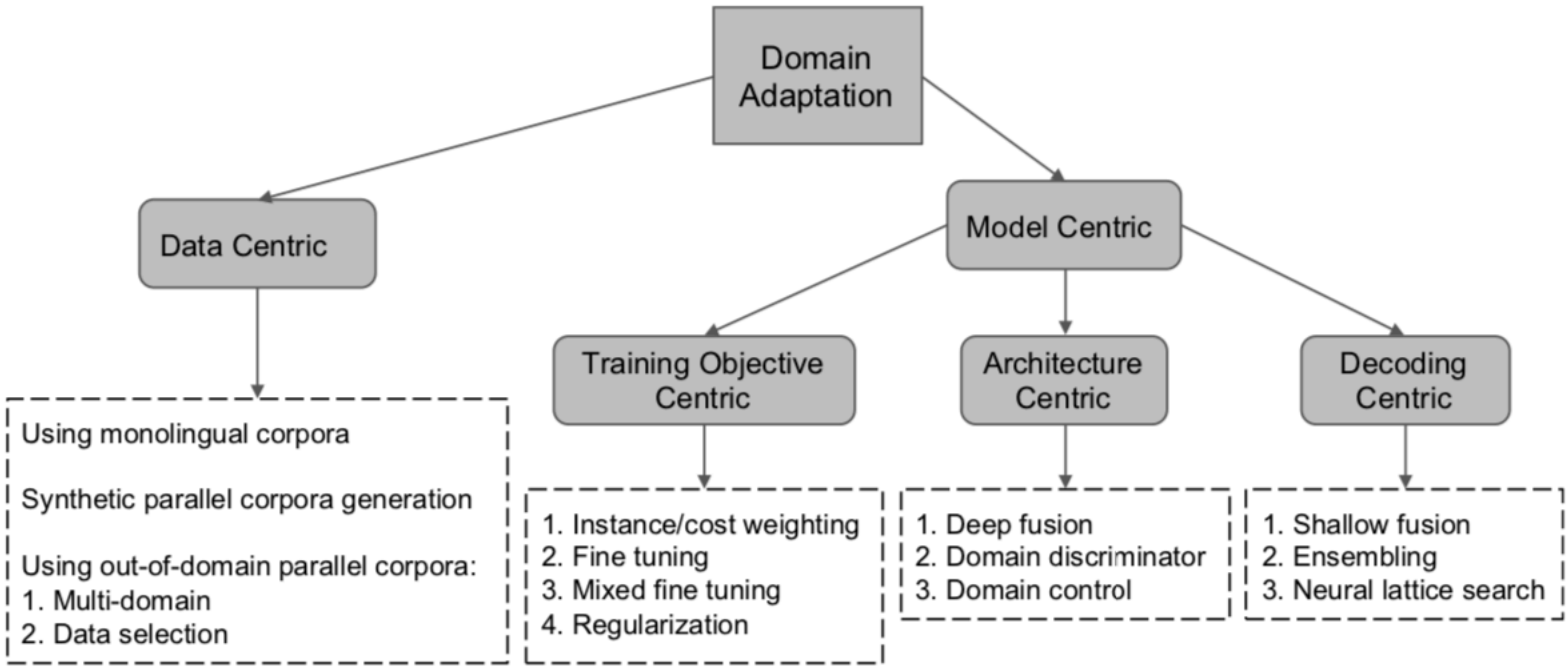
- Expensive to obtain training bitexts that are both large & relevant to test domain

Data Size

Relevance to test domain

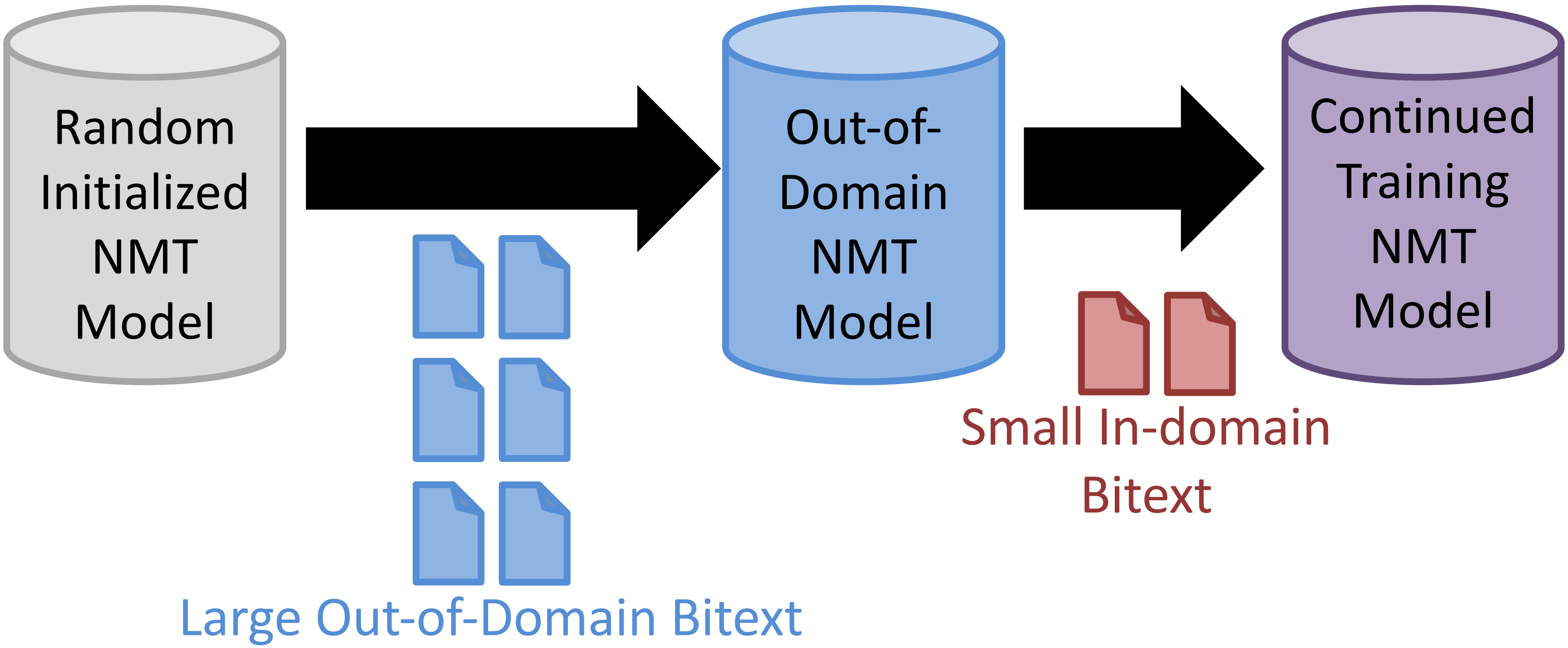
	Small	Large
Irrelevant		✔
Relevant	✔	✔✔

A taxonomy of domain adaptation methods for NMT



From: Chenhui Chu and Rui Wang, A Survey of Domain Adaptation for Neural Machine Translation, COLING 2018

Continued Training



Outline

1. Problem definition
2. Survey of adaptation methods
3. Error Analysis
   – S4 & Fluently Inadequate Translations
4. Promising Research Directions
   – CAT, new language adaptation, adaptation as window to understand NMT, etc.