Domain Adaptation for Neural Machine Translation

Chenhui Chu
Osaka University

Rui Wang
NICT

CCMT 2019 @ Nanchang
September 27th, 2019

Outline

1. Brief Introduction of Domain Adaptation (Wang)
2. Domain Adaptation for SMT (Wang)
3. Domain Adaptation for NMT (Chu)
4. Domain Adaptation in Specific Scenarios (Wang)
5. Datasets and Resources (Wang)
6. Future Directions (Wang)

Outline

1. Brief Introduction of Domain Adaptation
2. Domain Adaptation for SMT
3. Domain Adaptation for NMT
4. Domain Adaptation in Specific Scenarios
5. Datasets and Resources
6. Future Directions

Domain Adaptation

- Transfer learning: use of source domain *Ds* and source task *Ts* to improve the effect of target domain *Dt* and target task *Tt*
- The information of *Ds* and *Ts* is transferred to *Dt* and *Tt*
- Domain adaptation: a type of isomorphic transfer learning where *Ts* = *Tt*

Why do We Need Domain Adaptation? [Jiang+, 2007; Chang+ 2009]

- In-domain training data is small
- Different distributions
  - P(x): The distribution of training and testing data are different
  - P(y|x): With the same example, the label are different in different domains
- Unknown words
  - There are many unseen words in the new domain
- New Types
  - There are new types in the new domain (e.g., now predicting locations)

Domain Adaptation in Machine Translation:

- *Ds*: out-of-domain information (data, model etc.)
- *Dt*: in-domain information (data, model etc.)
- *Ts* = *Tt*: machine translation (statistical, neural etc.)

In this tutorial, we focus on empirical methods instead of mathematics and most of the references can be found at:

A Survey of Domain Adaptation for Neural Machine Translation, Chu and Wang, COLING-2018

Outline

Machine Translation

- Translation: to break the barrier between different cultures:

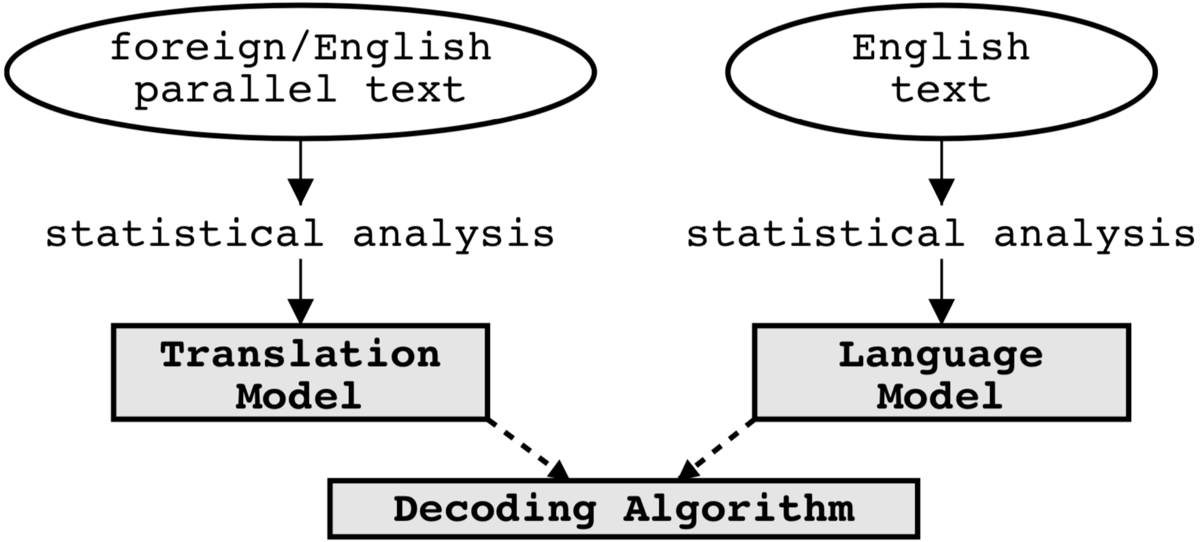
Type	Characteristics
Human Translation	Accurate but time-consuming
Machine Translation	Scalable but less accurate

- Machine Translation: a classic NLP/AI task
  - MT is a typical text generation task.
  - MT has standard evaluation Metrics.

- Reference Translation
  - the gunman was shot to death by the police .
- System Translations
  - the gunman was police kill .
  - wounded police jaya of
  - the gunman was shot dead by the police .
  - the gunman arrested by police kill .
  - the gunmen were killed .
  - the gunman was shot to death by the police .
  - gunmen were killed by police ?SUB>0 ?SUB>0
  - al by the police .
  - the ringer is killed by the police .
  - police killed the gunman .
- Matches
  - green = 4 gram match (good!)
  - red = word not matched (bad!)

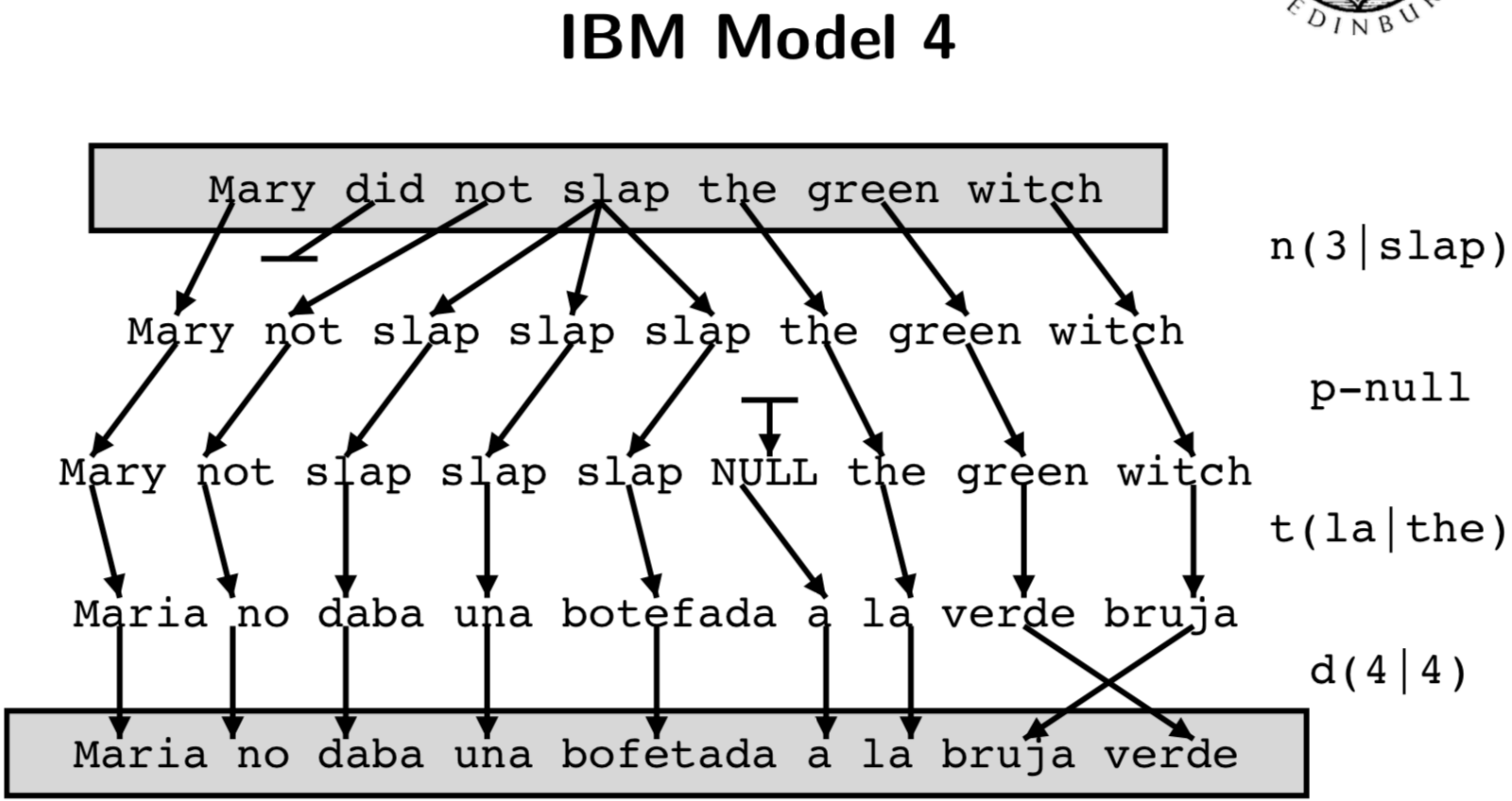
Statistical Machine Translation [Koehn, 2007]

Components: **Translation model**, **language model**, **decoder**



$$T(f) = \hat{e} = \operatorname{argmax}_e P(e)\, P(f|e)$$

Workflow of SMT



Toolkit: Moses [Koehn, 2007]

Moses
Software

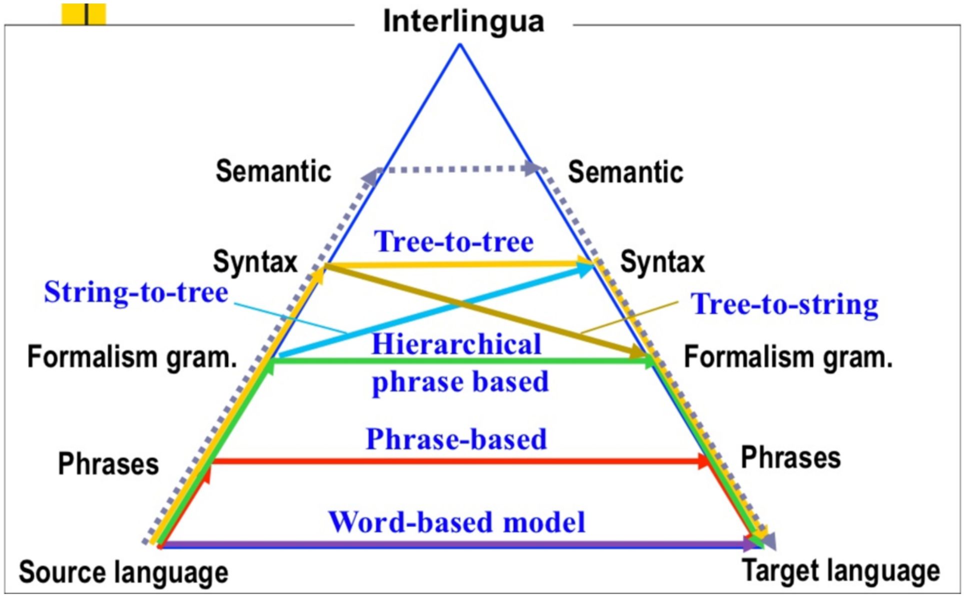
Moses is a free software, statistical machine translation engine that can be used to train statistical models of text translation from a source language to a target language. Moses then allows new source-language text to be decoded using these models to produce automatic translations in the target language. Wikipedia

Operating system: Windows, Linux, macOS

Stable release: 4.0 / October 5, 2017; 22 months ago

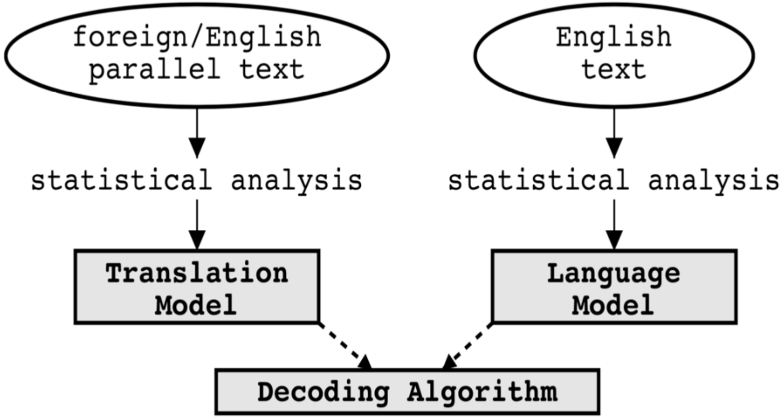
License: LGPL

Written in: C++, Perl

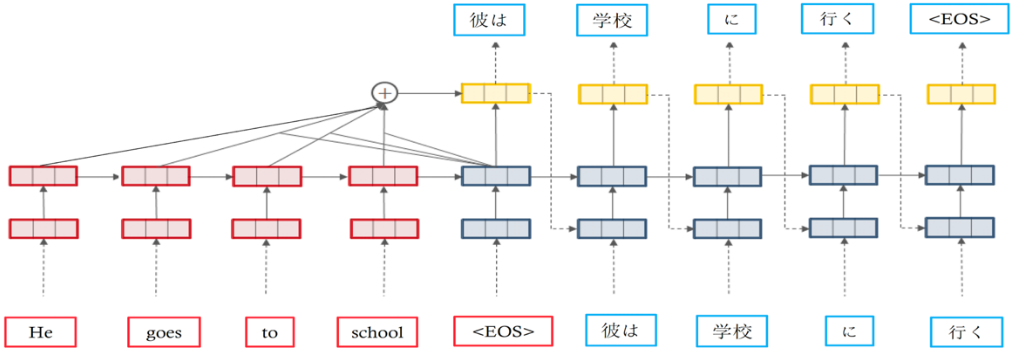


SMT vs NMT

Components: **Translation model**, **language model**, **decoder**



SMT



NMT

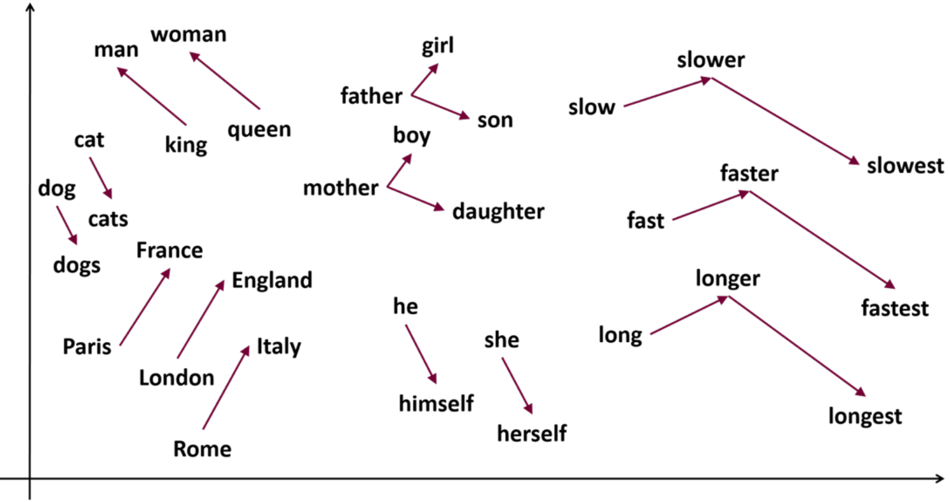
Phrase Table (Translation Model) in SMT

```
> grep '| in europe |' model/phrase-table | sort -nrk 7 -t\| | head
in europa ||| in europe ||| 0.829007 0.207955 0.801493 0.492402
europas ||| in europe ||| 0.0251019 0.066211 0.0342506 0.0079563
in der europaeischen union ||| in europe ||| 0.018451 0.00100126 0.0319584 0.0196869
in europa , ||| in europe ||| 0.011371 0.207955 0.207843 0.492402
europaeischen ||| in europe ||| 0.00686548 0.0754338 0.000863791 0.046128
im europaeischen ||| in europe ||| 0.00579275 0.00914601 0.0241287 0.0162482
fuer europa ||| in europe ||| 0.00493456 0.0132369 0.0372168 0.0511473
in europa zu ||| in europe ||| 0.00429092 0.207955 0.714286 0.492402
an europa ||| in europe ||| 0.00386183 0.0114416 0.352941 0.118441
der europaeischen ||| in europe ||| 0.00343274 0.00141532 0.00099583 0.000512159
```

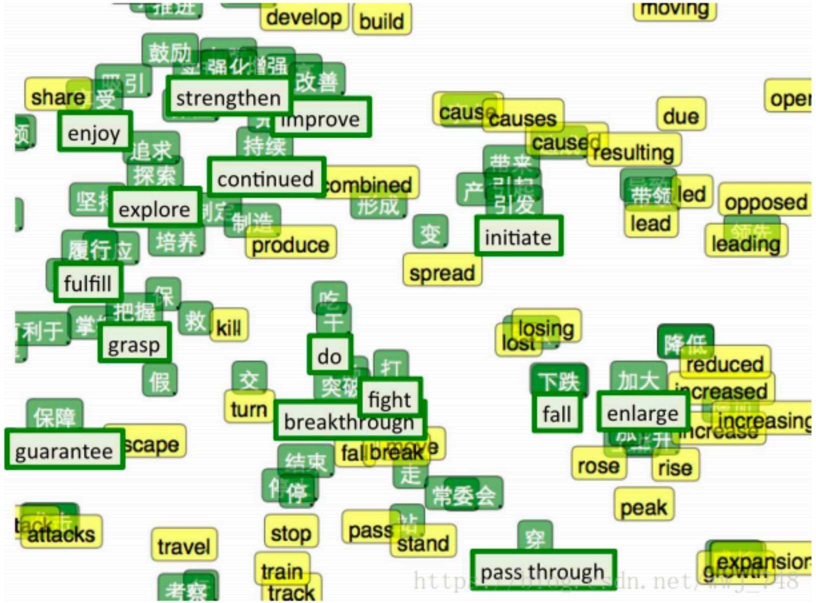
four different phrase translation scores are computed:

1. inverse phrase translation probability $\varphi(f|e)$
2. inverse lexical weighting $lex(f|e)$
3. direct phrase translation probability $\varphi(e|f)$
4. direct lexical weighting $lex(e|f)$

Bilingual Word Embedding



Monolingual Word Embedding



Bilingual Word Embedding

Language Model

- A language model (LM) is a model that assigns a probability to a sentence.
- *N*-gram LM

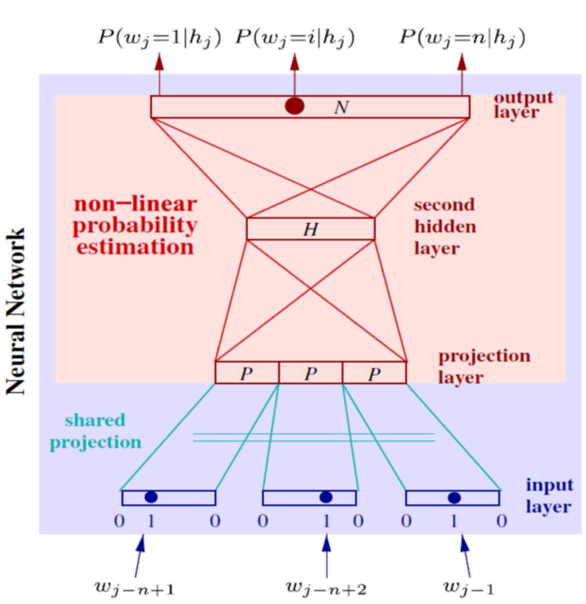
In an *n*-gram model, the probability $P(w_1, \ldots, w_m)$ of observing the sentence $w_1, \ldots, w_m$ is approximated as

$$P(w_1, \ldots, w_m) = \prod_{i=1}^{m} P(w_i \mid w_1, \ldots, w_{i-1}) \approx \prod_{i=1}^{m} P(w_i \mid w_{i-(n-1)}, \ldots, w_{i-1})$$

Format (example)

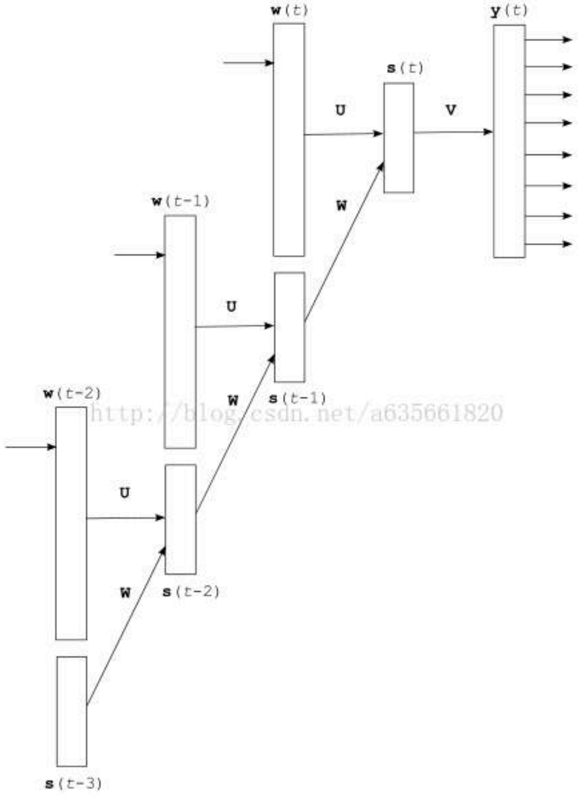
2-grams		3-grams	
-1.7037368	<s> I	-1.4910358	<s> I am
-3.1241505	a boy	-1.1888235	I am a
-1.9892355	am a	-0.6548149	a boy .
-1.0562452	boy .	-1.1425415	. </s> 0

Neural Network Language Model



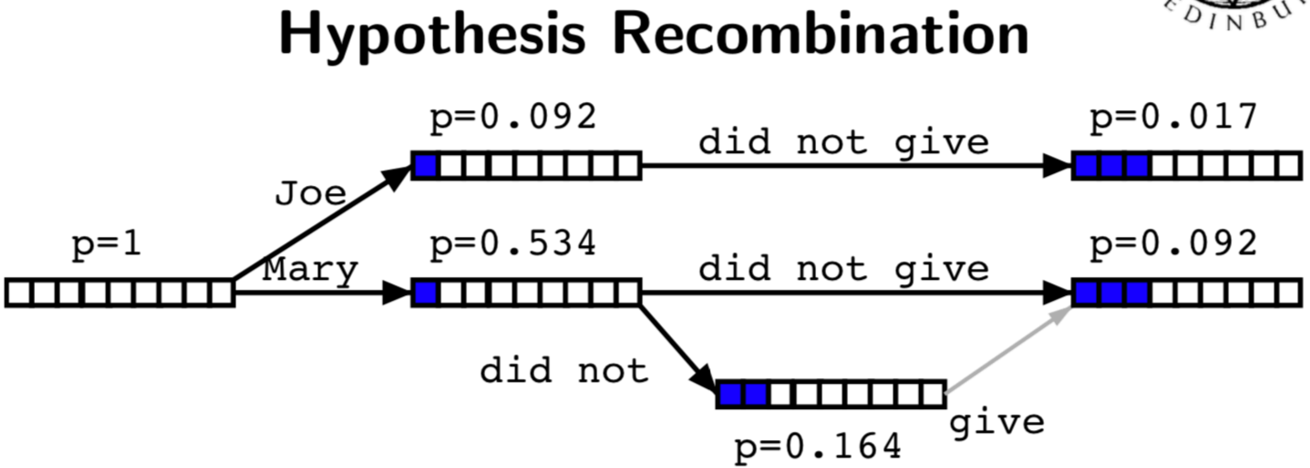
$$P(w_i|h_i) = \begin{cases} \dfrac{P_c(w_i|h_i)}{1 - P_c(o|h_i)} P_s(h_i), & if\ w_i \in shortlist \\ P_b(w_i|h_i), & otherwise \end{cases}$$

Continuous-space LM (CSLM) or NNLM [Schwenk, 2010]



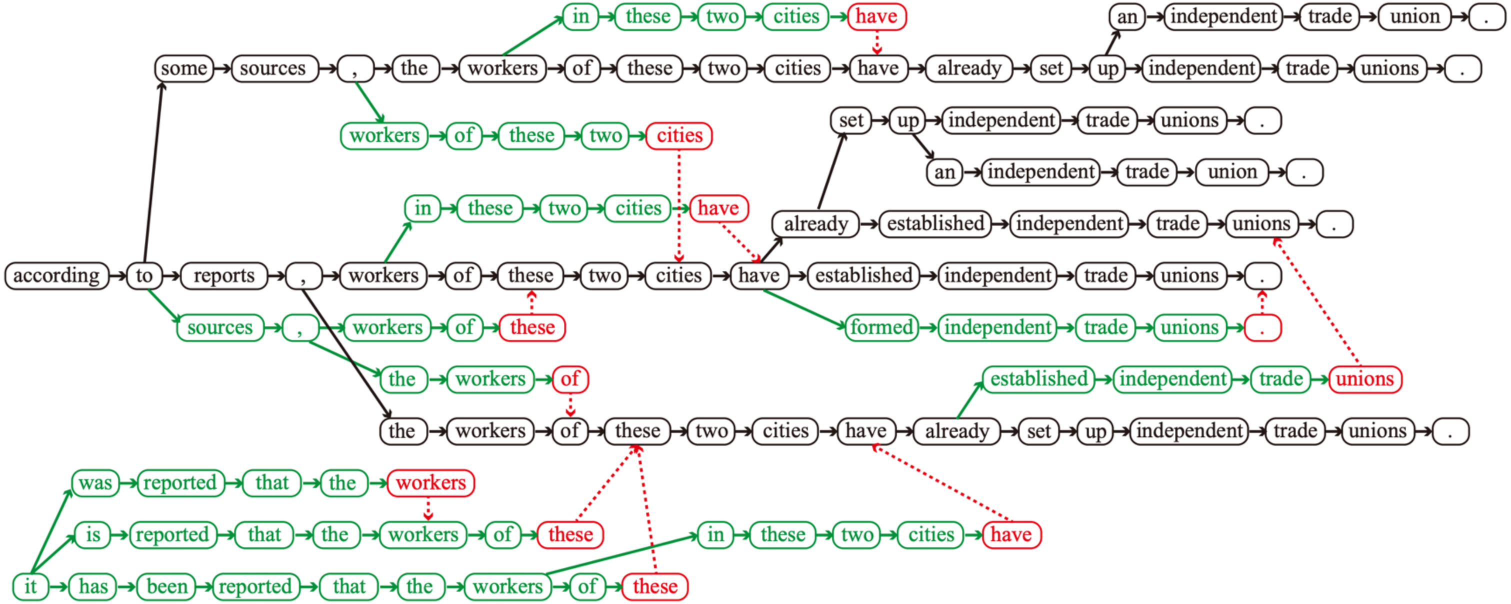
RNNLM [Mikolov, 2012]

Decoding in SMT



- Recombined hypotheses do *not* have to *match completely*
- No matter what is added, weaker path can be dropped, if:
  - *last two English words* match (matters for language model)
  - *foreign word coverage* vectors match (effects future path)

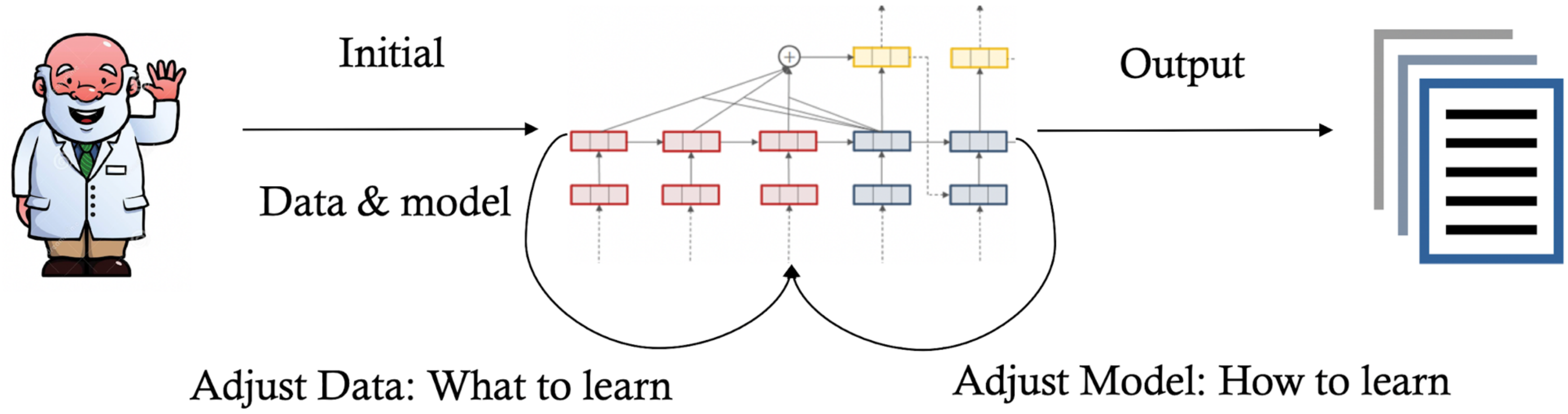
Decoding in NMT



[Zhang and Wang et al., 2010]

Domain Adaptation for Machine Translation

1. Data Centric
2. Model Centric



Outline

1. Brief Introduction of Domain Adaptation
2. Domain Adaptation for SMT
    a. Data Centric
    b. Model Centric
3. Domain Adaptation for NMT
4. Domain Adaptation in Specific Scenarios
5. Datasets and Resources
6. Future Directions

Data Selection/Generation for SMT

- Sufficient parallel corpora: select parallel sentences from out-of-domain parallel sentences by some criteria
  - Cross-entropy by LM [Moore+, 2010; Axelrod+ 2011; Duh+, 2013]
  - EM training algorithm [Hoang+, 2014]
  - Convolutional neural network classifier [Chen+, 2016]
- Insufficient parallel corpora: generate pseudo-parallel sentences by some criteria
  - Information retrieval [Utiyama+, 2003]
  - Bilingual word embeddings [Marie and Fujita, 2017]
  - Generate parallel phrase pairs [Chu+, 2015; Wang+, 2016]

Example1: Cross-Entropy based Data Selection

Cross Entropy: The cross entropy for the distributions *p* and *q* over a given set is defined as follows:

$$H(p,q)=E_p[-\log q]$$

Monolingual sentence selection criteria [Moore+ 2010]

$$HI(s) - HO(s)$$

Bilingual sentence selection criteria [Axelrod+ 2011]

$$[HI\_src(s)-HO\_src(s)]+[HI\_tgt(s)-HO\_tgt(s)]$$

Example 2: Phrase Generation [Wang+ 2016]

Phrase is a small and more fine grained unit for data selection

- Two phrases 'would like to learn' and 'Chinese as second language' are in the in-domain PT. In decoding, these two phrases may be connected together as 'would like to learn Chinese as second language"
- The phrases 'would like to learn Chinese' or 'learn Chinese as second language' may be used as the new generated *n*-gram LM or phrases in phrase-table

Outline

1. Brief Introduction of Domain Adaptation
2. Domain Adaptation for SMT
   a. Data Centric
   b. Model Centric
3. Domain Adaptation for NMT
4. Domain Adaptation in Specific Scenarios
5. Datasets and Resources
6. Future Directions

- Model level interpolation [Foster+, 2007; Bisazza+, 2011; Niehues+, 2012; Sennrich+, 2013; Durrani+, 2015; Imamura+, 2016]
  - Several SMT models, such as LMs, translation, and reordering models, corresponding to each corpus, are individually trained
  - These models are then combined to achieve the best performance
- Instance level interpolation [Jiang+, (2007)]
  - Firstly score each instance/domain by using rules or statistical methods as a weight
  - Then train SMT models by giving each instance/domain the weight

Model Level Interpolation

- Interpolation [Foster+, 2007]
    - Split the corpus into different components, according to some criterion
    - Train a model on each corpus component
    - Weight each model according to its fit with the test domain
    - Combine weighted component models into a single global model
- Fill-up [Bisazza et al., 2011]
    - First, separate translation models are built from in-domain and background data
    - The background table is merged with the in-domain table by adding only new phrase pairs that do not appear in the in-domain table

Instance Level Interpolation [Jiang+, 2007]

- Defined the domain adaptation problem in NLP as:
  - $p_s(x, y)$ and $p_t(x, y)$: distributions for the source and the target domains
  - Use $p_s(x, y)$ to approximate $p_t(x, y)$
- In MT, simplify domain adaptation as:

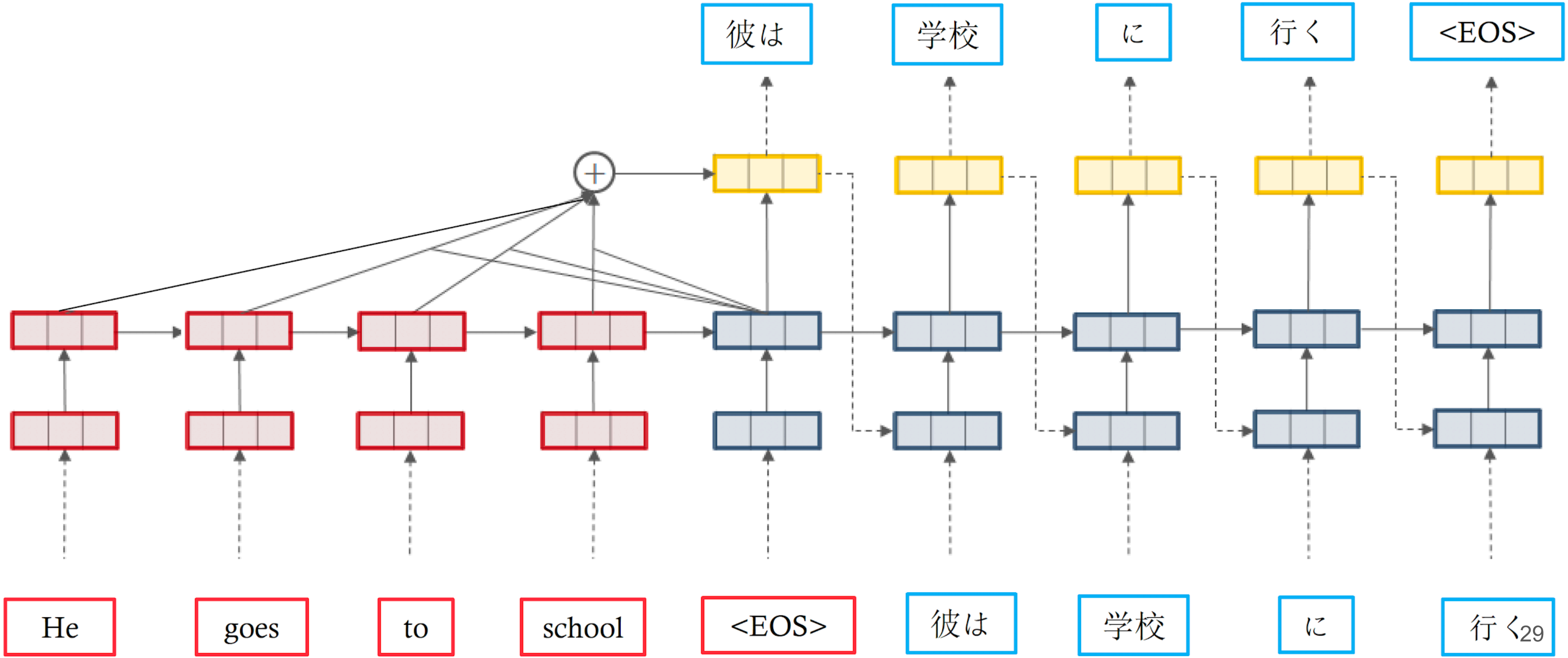
$$J_{dw} = \lambda_{in} \sum_{(\mathbf{x},\mathbf{y}) \in \mathcal{D}_{in}} logp(\mathbf{y}|\mathbf{x}) + \sum_{(\mathbf{x}',\mathbf{y}') \in \mathcal{D}_{out}} logp(\mathbf{y}'|\mathbf{x}').$$

In-domain weight

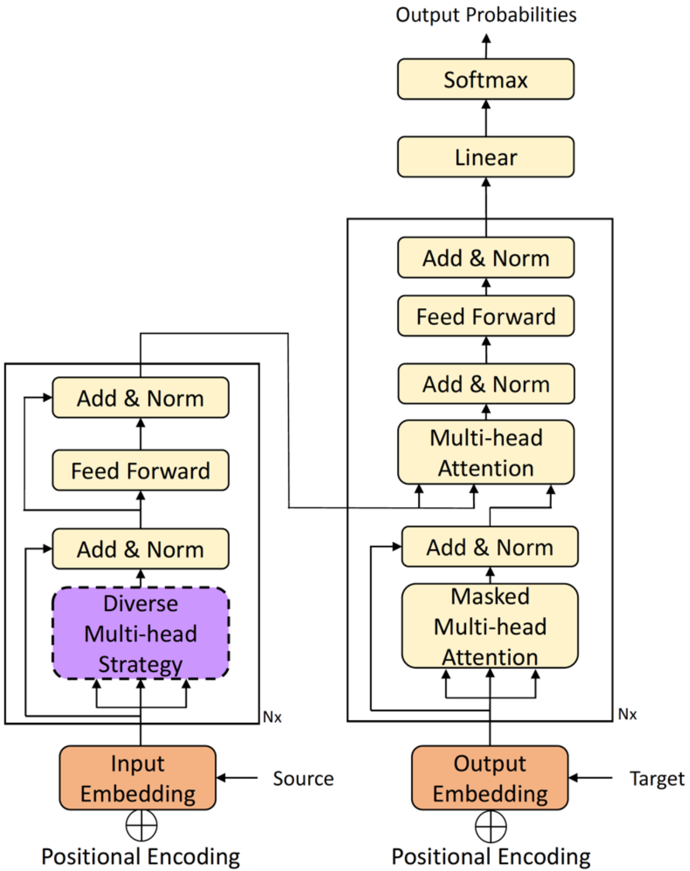
Outline

1. Brief Introduction of Domain Adaptation
2. Domain Adaptation for SMT
3. Domain Adaptation for NMT
4. Domain Adaptation in Specific Scenarios
5. Datasets and Resources
6. Future Directions

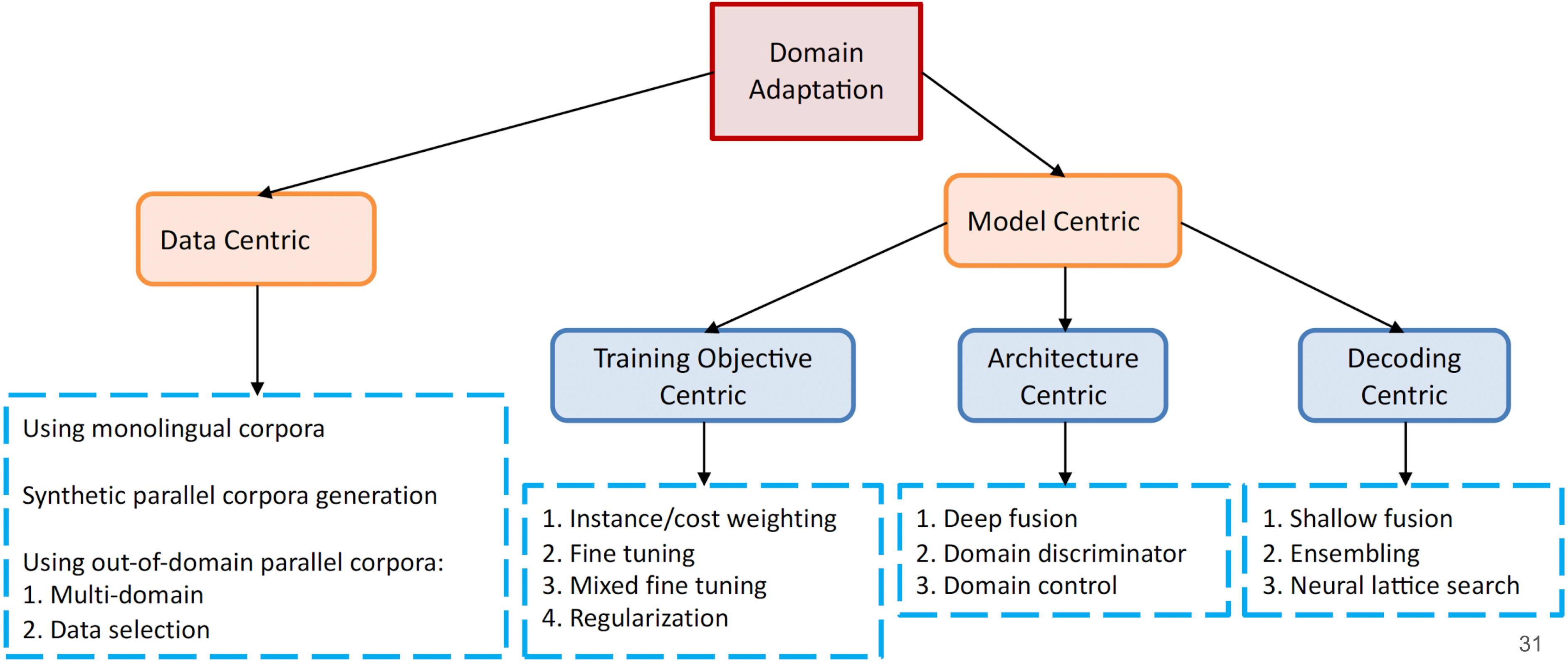
RNN based NMT [Bahdanau+ 2015]



Self-Attention Based NMT [Vaswani+ 2017]

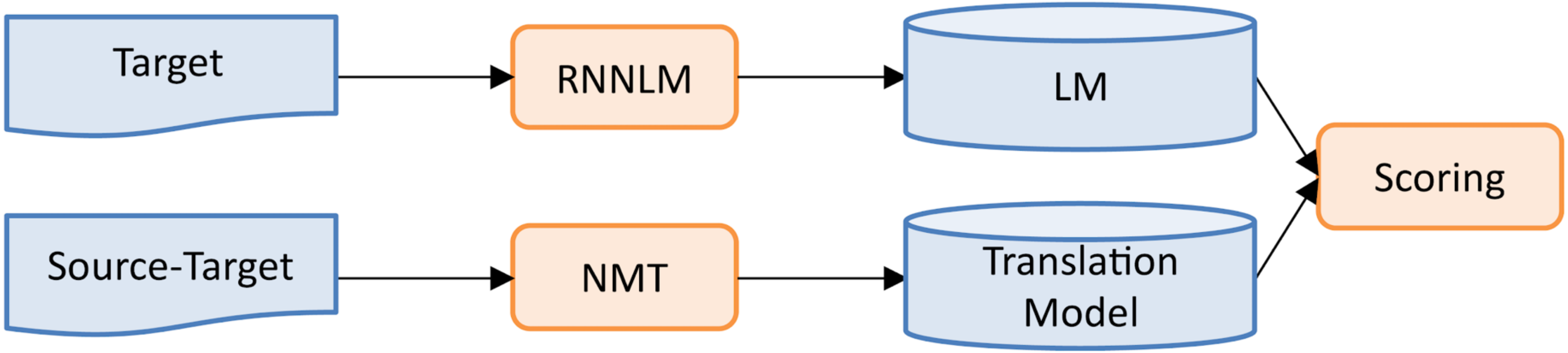


Overview of Domain Adaptation for NMT [Chu+ 2018]

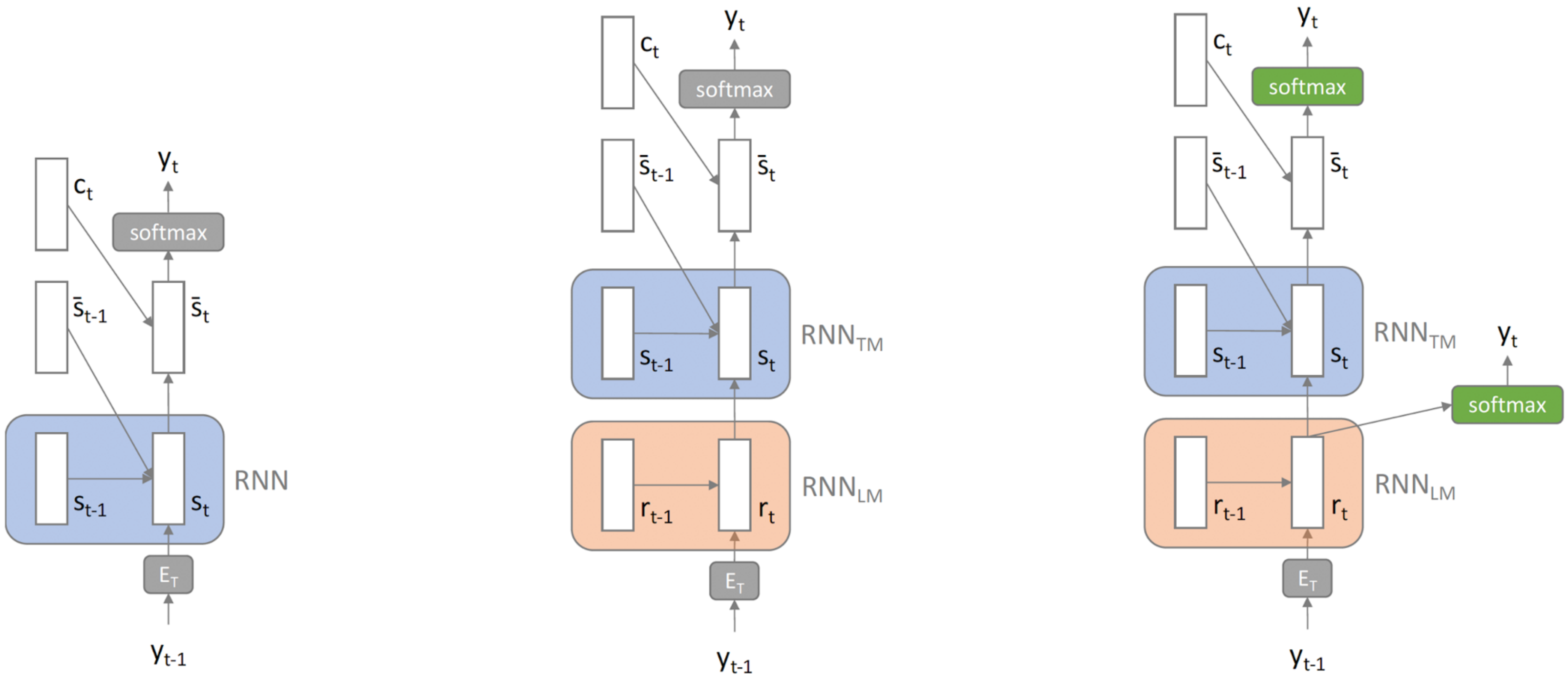


Outline

Target-Side RNNLM Fusion [Gulcehre+ 2015]



Target-Side Multi-task Learning [Domhan+ 2017]



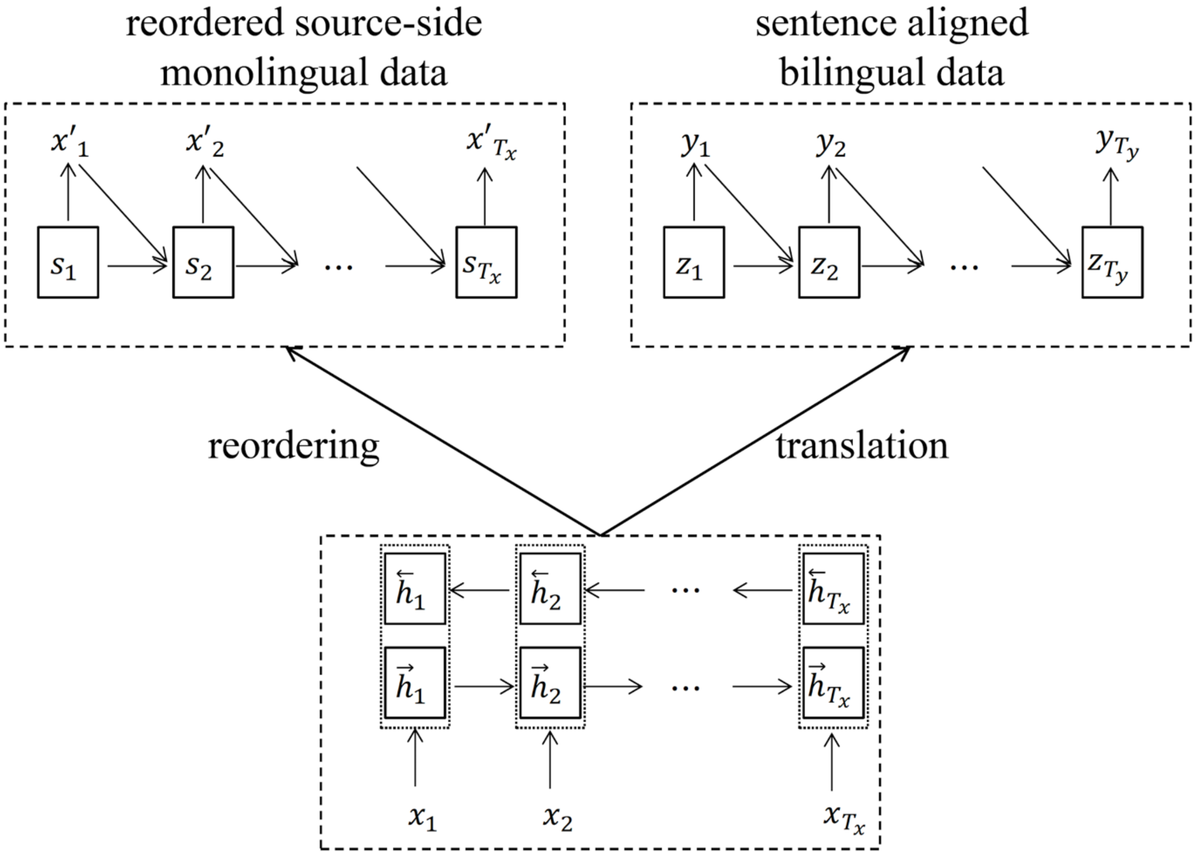
(a) baseline (b) +LML (c) +LML +MTL

Results of Multi-task Learning [Domhan+ 2017]

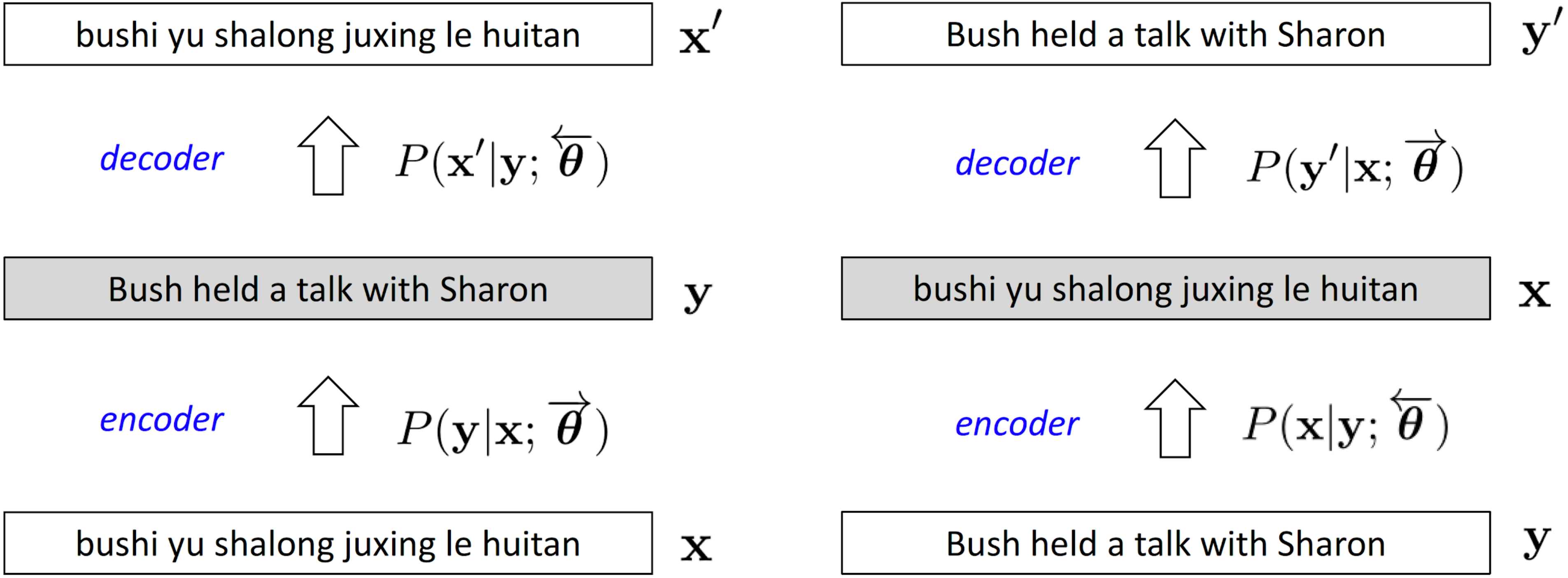
System	Data	EN→DE	FR→EN	CS→EN
baseline		20.3 39.9 63.0	21.7 27.5 59.1	17.0 24.4 65.2
+ LML		20.4 39.8 63.1	21.3 27.2 59.8	16.9 24.4 65.4
+ LML + MTL	+ mono	21.4 40.8 61.4	22.3 27.7 58.3	17.2 24.7 64.3
Sennrich et al. (2016)	+ synthetic	24.4 43.4 56.4	27.4 31.5 52.1	21.2 27.5 59.4
ensemble baseline		22.2 41.6 60.6	23.9 29.1 56.4	18.3 25.5 63.0
+ LML		22.4 41.8 60.9	23.5 28.7 57.2	18.3 25.6 63.4
+ LML + MTL	+ mono	23.6 42.8 58.9	24.2 29.2 55.9	18.8 25.9 62.2
ensemble Sennrich et al. (2016)	+ synthetic	25.7 44.6 55.0	29.1 32.6 50.3	22.5 28.4 57.8

Table 1: BLEU/METEOR/TER scores on test sets for different language pairs. For BLEU and METEOR higher is better. For TER lower is better.

Source-Side Multi-task Learning [Zhang+ 2016]



Both Source and Target-Side with Autoencoder [Cheng+ 2016]



(a) (b)

Results of Autoencoder [Cheng+ 2016]

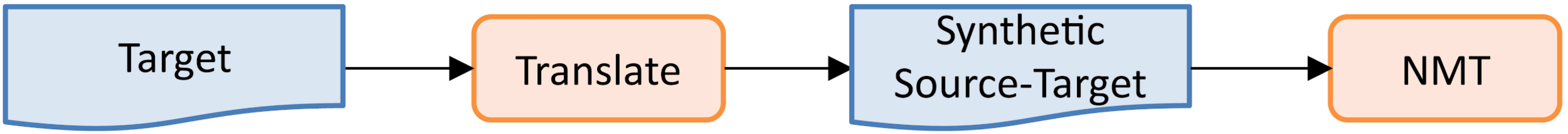
Method	Training Data CE	C	E	Direction	NIST06	NIST02	NIST03	NIST04	NIST05
Sennrich et al. (2015)	√	×	√	C →E	34.10	36.95	36.80	37.99	35.33
	√	√	×	E →C	19.85	28.83	20.61	20.54	19.17
this work	√	×	√	C →E	35.61**	38.78**	38.32**	38.49*	36.45**
				E →C	17.59	23.99	18.95	18.85	17.91
	√	√	×	C →E	35.01**	38.20**	37.99**	38.16	36.07**
				E →C	21.12**	29.52**	20.49	21.59**	19.97**

Either E or C can be used on both sides for either C->E or E->C

Outline

1. Brief Introduction of Domain Adaptation
2. Domain Adaptation for SMT
3. Domain Adaptation for NMT
    a. Data Centric
        i. Using Monolingual Corpora
        ii. Synthetic Parallel Corpora Generation
        iii. Using Out-of-Domain Parallel Corpora
    b. Model Centric
4. Domain Adaptation in Specific Scenarios
5. Datasets and Resources
6. Future Directions

Synthetic Parallel Corpora Generation [Sennrich+ 2016]



Bi-Directional Parallel Corpora Generation [Niu+ 2018]

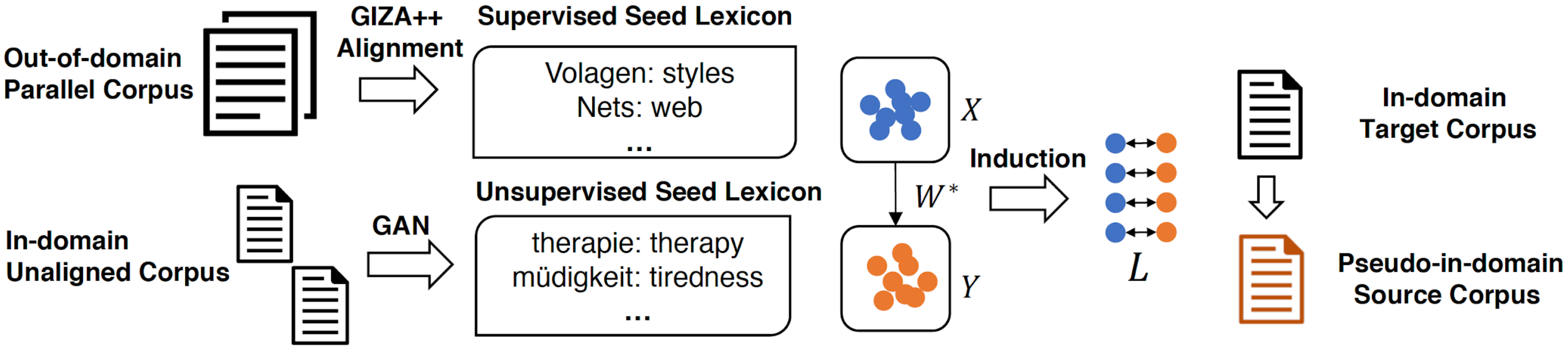
Target better

Both better

ID	Training Data	TL→EN	EN→TL	SW→EN	EN→SW	DE→EN	EN→DE
U-1	L1→L2	31.99	31.28	32.60	39.98	29.51	23.01
U-2	L1→L2 + L1⋆→L2	**24.21**	**29.68**	**25.84**	**38.29**	**33.20**	**25.41**
U-3	L1→L2 + L1→L2⋆	22.13	27.14	24.89	36.53	30.89	23.72
U-4	L1→L2 + L1⋆→L2 + L1→L2⋆	23.38	29.31	25.33	37.46	33.01	25.05
	L1=EN	L2=TL		L2=SW		L2=DE	
B-1	L1↔L2	32.72	31.66	33.59	39.12	28.84	22.45
B-2	L1↔L2 + L1⋆↔L2	32.90	**32.33**	33.70	**39.68**	29.17	**24.45**
B-3	L1↔L2 + L2⋆↔L1	32.71	31.10	33.70	39.17	**31.71**	21.71
B-4	L1↔L2 + L1⋆↔L2 + L2⋆↔L1	**33.25**	**32.46**	**34.23**	38.97	30.43	22.54
B-5	L1↔L2 + L1⋆→L2 + L2⋆→L1	**33.41**	**33.21**	**34.11**	**40.24**	**31.83**	**24.61**
B-5*	L1↔L2 + L1⋆→L2 + L2⋆→L1	33.79	32.97	34.15	40.61	31.94	24.45
B-6*	L1↔L2 + <u>L1⋆</u>→L2 + <u>L2⋆</u>→L1	**34.50**	**33.73**	**34.88**	**41.53**	**32.49**	**25.20**

Table 2: BLEU scores for uni-directional models (U-*) and bi-directional NMT models (B-*) trained on different combinations of real and synthetic parallel data. Models in B-5* are fine-tuned from base models in B-1. Best models in B-6* are fine-tuned from precedent models in B-5* and underscored synthetic data is re-decoded using precedent models. Scores with largest improvement within each zone are highlighted.

Synthetic Data by Lexicon Induction [Hu+ 2019]



Results of Synthetic Data by Lexicon Induction [Hu+ 2019]

	Medical	Subtitles	Law	Koran
Unadapted	7.43	5.49	4.10	2.52
Copy	13.28	6.68	5.32	3.22
BT	18.51	11.25	11.55	**8.18**
DALI-U	20.44	9.53	8.63	4.90
DALI-S	19.03	9.80	8.64	4.91
DALI-U+BT	**24.34**	**13.35**	**13.74**	8.11
DALI-GIZA++	28.39	9.37	11.45	8.09
In-domain	46.19	27.29	40.52	19.40

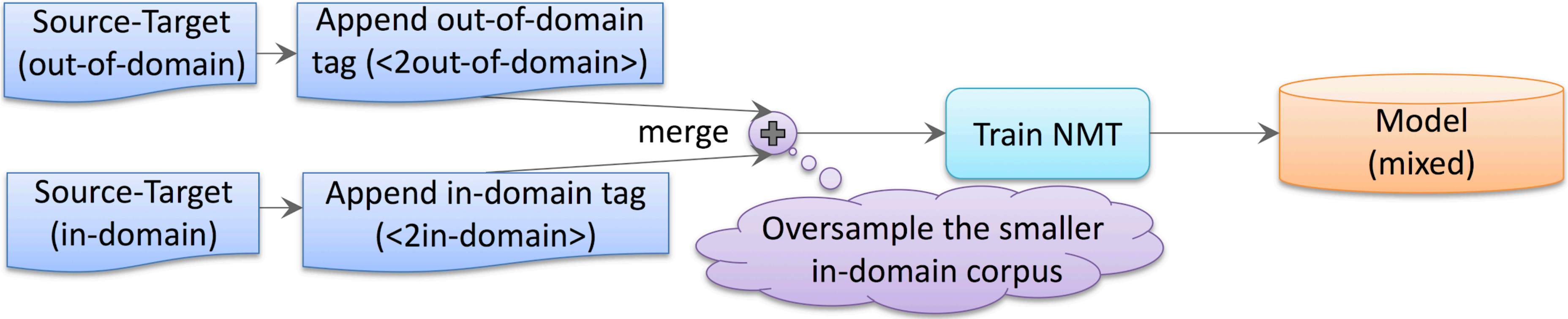
Upper bound

Table 3: Comparison among different methods on adapting NMT from IT to {Medical, Subtitles, Law, Koran} domains, along with two oracle results

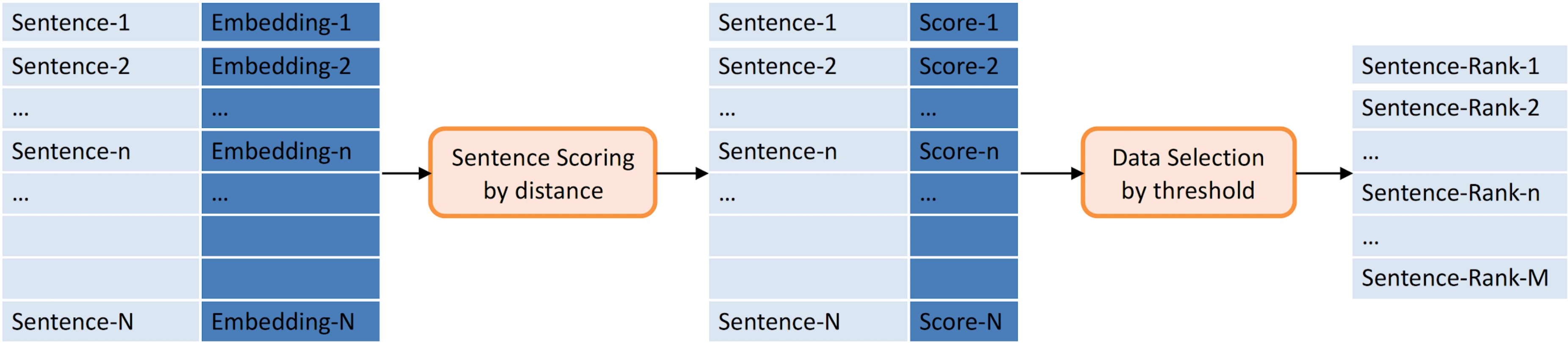
Outline

1. Brief Introduction of Domain Adaptation
2. Domain Adaptation for SMT
3. Domain Adaptation for NMT
    a. Data Centric
        i. Using Monolingual Corpora
        ii. Synthetic Parallel Corpora Generation
        iii. Using Out-of-Domain Parallel Corpora
    b. Model Centric
4. Domain Adaptation in Specific Scenarios
5. Datasets and Resources
6. Future Directions

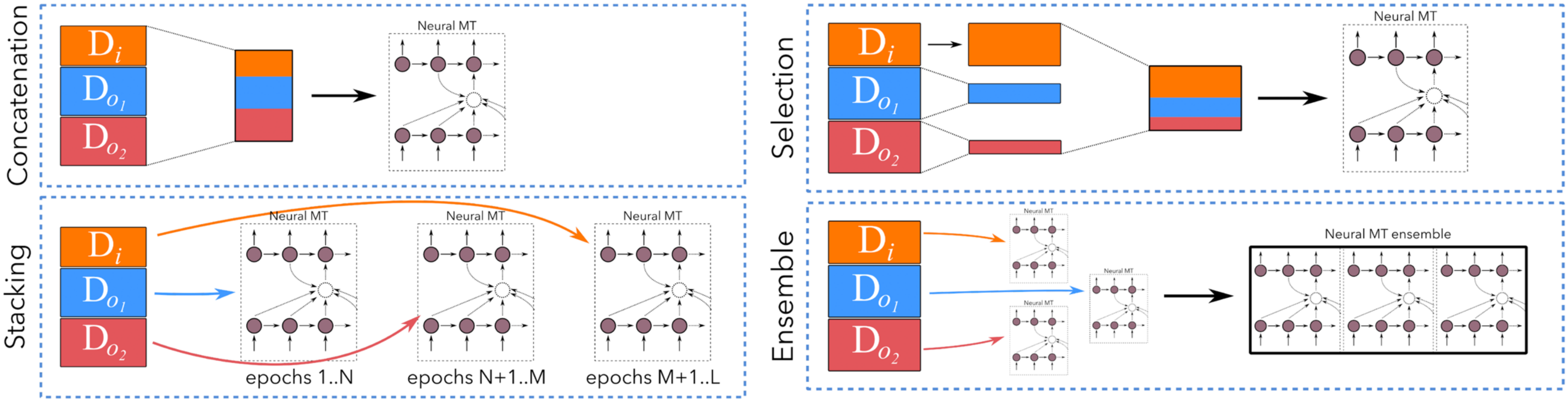
Multi-Domain [Kobus+ 2016]



Data Selection [Wang+ 2017]



Different Multi-Domain Approaches [Sajjad+ 2017]



Results of Multi-Domain Approaches [Sajjad+ 2017]

Fine-tuning

		Arabic-English	
	ALL	OD→TED	UN→OPUS→TED
tst13	36.1	37.9	36.8
tst14	30.2	32.1	31.2
avg.	33.2	35.0	34.0
		German-English	
	ALL	OD→TED	EP→CC→TED
tst13	35.7	38.1	36.8
tst14	30.8	32.8	31.7
avg.	33.3	35.4	34.3

Table 4: Stacking versus concatenation

	Arabic-English		German-English	
	ALL	Selected	ALL	Selected
tst13	36.1	32.7	35.7	34.1
tst14	30.2	27.8	30.8	29.9
avg.	33.2	30.3	33.3	32.0

	Arabic-English			
	OPUS	ALL	ENS_b	ENS_w
tst13	32.2	36.1	31.9	34.3
tst14	27.3	30.2	25.8	28.6
avg.	29.7	33.2	28.9	31.5

Table 6: Comparing results of balanced ensemble (ENS_b) and weighted ensemble (ENS_w) with the best individual model and the concatenated model

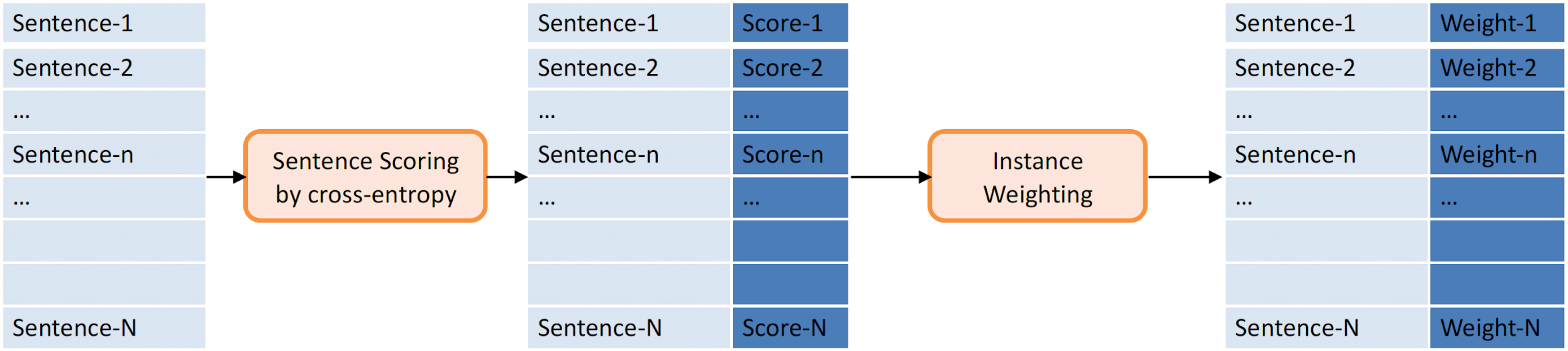
Outline

1. Brief Introduction of Domain Adaptation
2. Domain Adaptation for SMT
3. Domain Adaptation for NMT
    a. Data Centric
    b. Model Centric
        i. Training Objective Centric
        ii. Architecture Centric
        iii. Decoding Centric
4. Domain Adaptation in Specific Scenarios
5. Datasets and Resources
6. Future Directions

Outline

1. Brief Introduction of Domain Adaptation
2. Domain Adaptation for SMT
3. Domain Adaptation for NMT
    a. Data Centric
    b. Model Centric
        i. Training Objective Centric: Sentence Weighting
        ii. Architecture Centric
        iii. Decoding Centric
4. Domain Adaptation in Specific Scenarios
5. Datasets and Resources
6. Future Directions

Sentence Weighting [Wang+ 2017]



Sentence Selection and Weighting [Wang+ 2018]

sentence selection is a special case of sentence weighting, i.e., the sentences with low-weights are cut off

the weights are set to be the same

balance the ratio of the in-domain and out-of-domain data

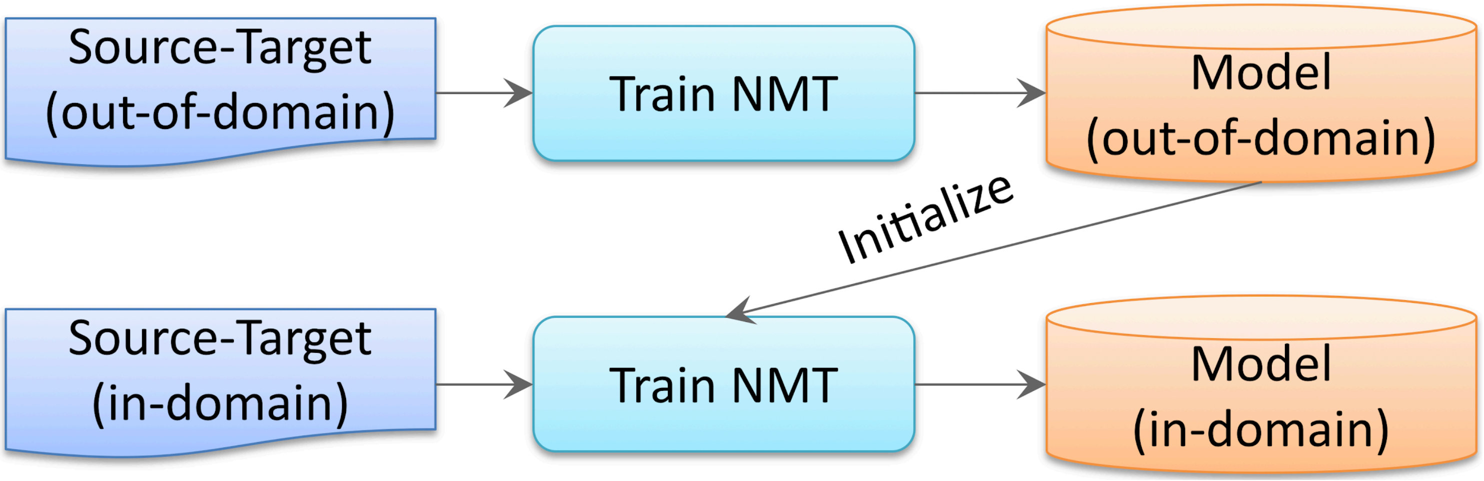
Results of sentence Selection and Weighting [Wang+ 2018]

IWSLT EN-FR	dev10	test10	test11
in	27.66	32.11	35.22
out	24.93	29.60	32.27
$in + out$	25.14	29.94	33.50
ensemble ($in + out$)	28.48	33.63	37.67
sampler	**28.67**	**34.12**	38.08
Kobus [54]	27.87	33.81	37.44
Axelrod [35]	27.85	34.03	**38.30**
sentence selection ($\delta_{\mathbf{fe}}$)	29.38+	35.57++	39.20++
sentence weighting	29.14+	34.80+	38.73
batch weighting	29.81++	35.54++	39.48++
sentence scoring+sentence weighting	29.97++	35.64++	40.17++
sentence selection+batch weighting	**30.17++**	**36.03++**	**40.59++**

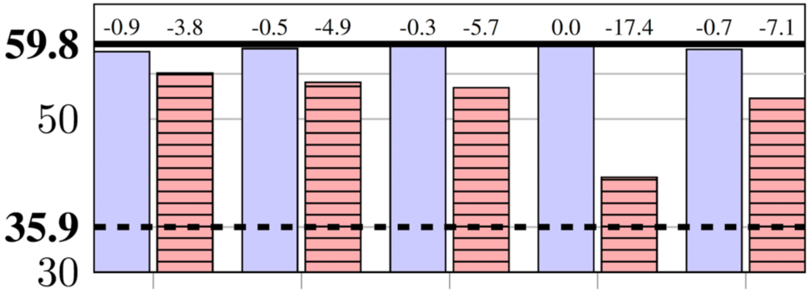
Outline

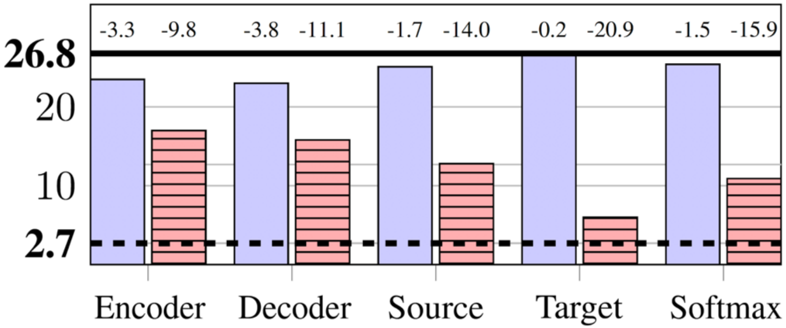
1. Brief Introduction of Domain Adaptation
2. Domain Adaptation for SMT
3. Domain Adaptation for NMT
    a. Data Centric
    b. Model Centric
        i. Training Objective Centric: Fine Tuning
        ii. Architecture Centric
        iii. Decoding Centric
4. Domain Adaptation in Specific Scenarios
5. Datasets and Resources
6. Future Directions

Fine Tuning [Luong+ 2015; Sennrich+ 2016; Servan+ 2016; Freitag+ 2016]

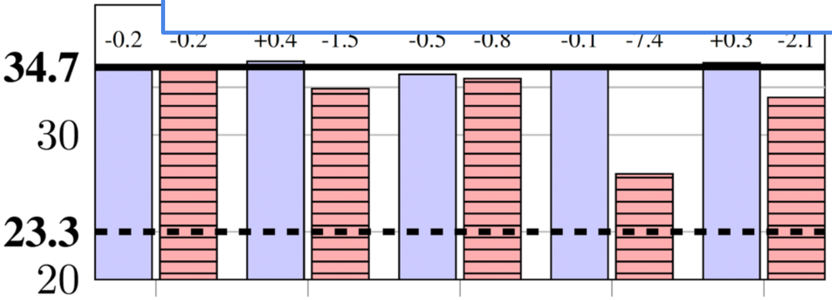


Effects of Components in Fine Tuning [Thompson+ 2018]



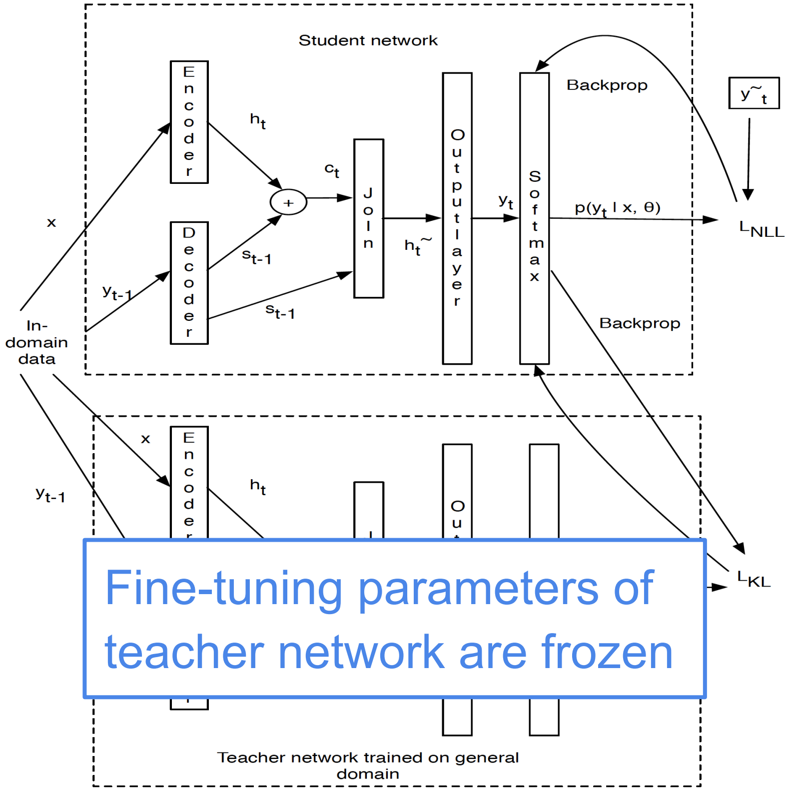


Any single component has little impact on the performance

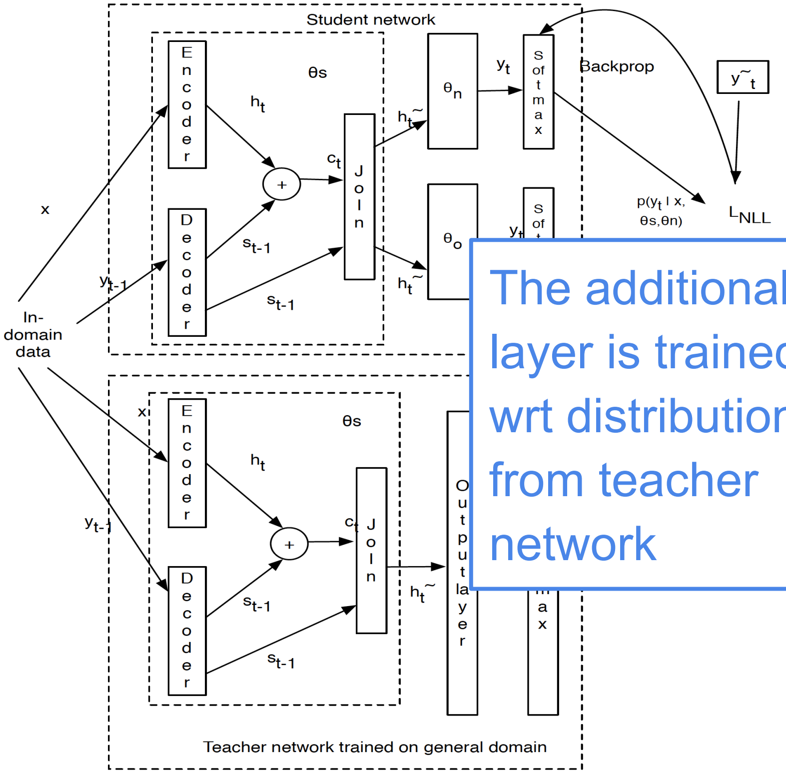


(b) Results on WIPO Ru–En

Prevent Out-of-domain Translation Degradation [Dakwale+ 2017]

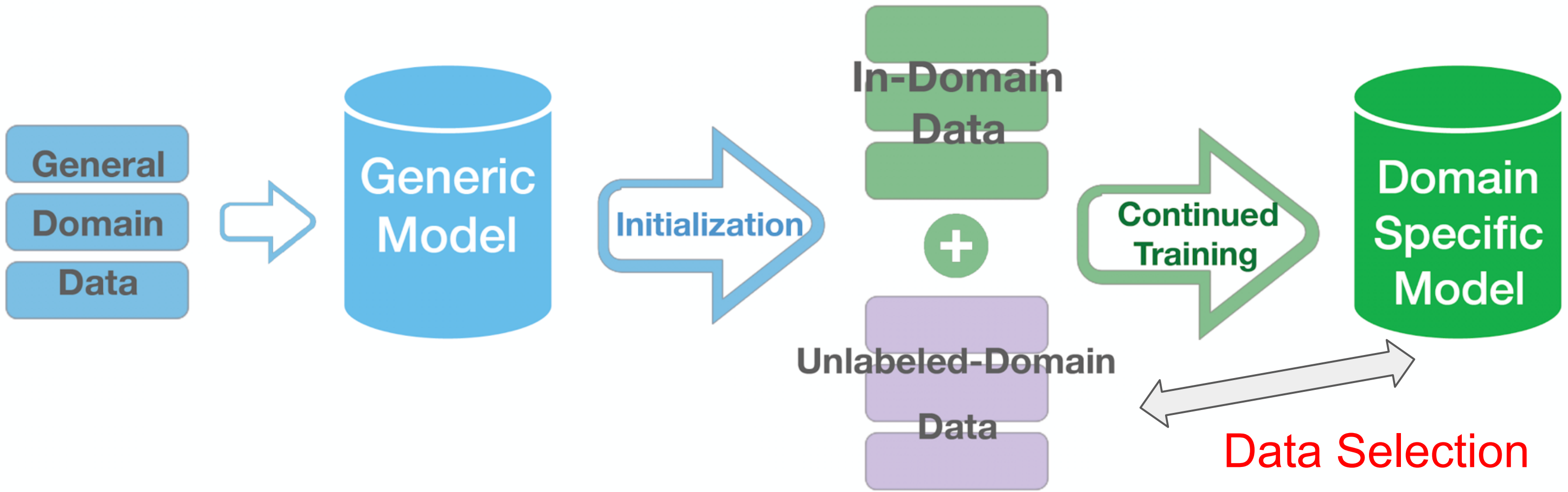


Fine-tuning parameters of teacher network are frozen



The additional layer is trained wrt distribution from teacher network

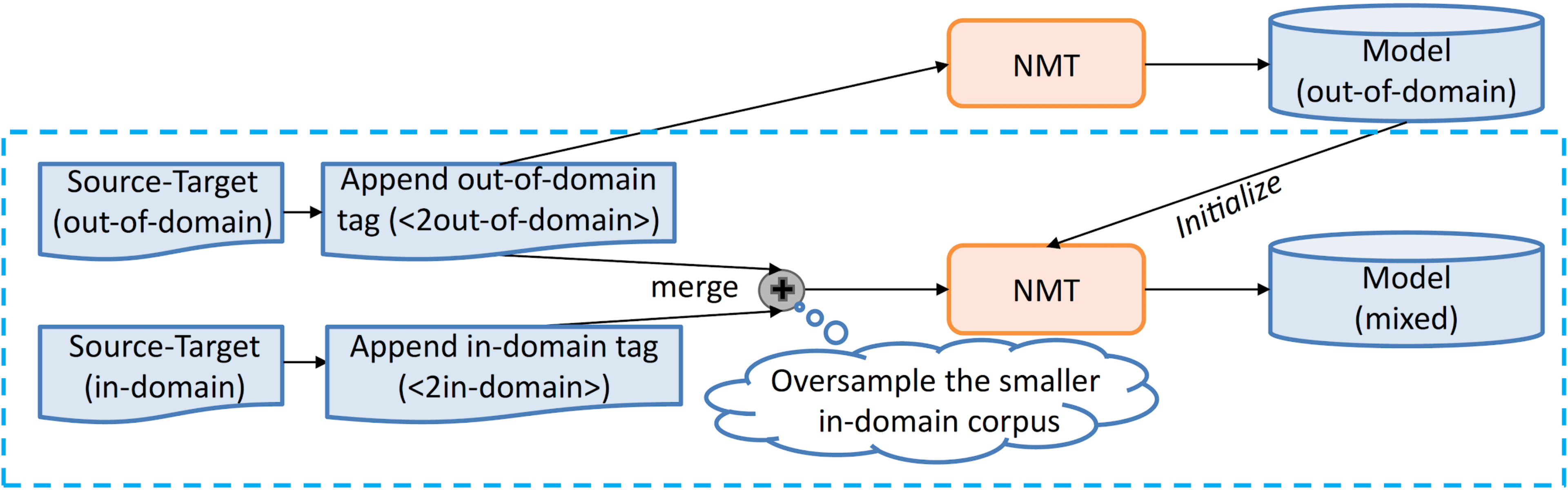
Curriculum Learning [Zhang+ 2019]



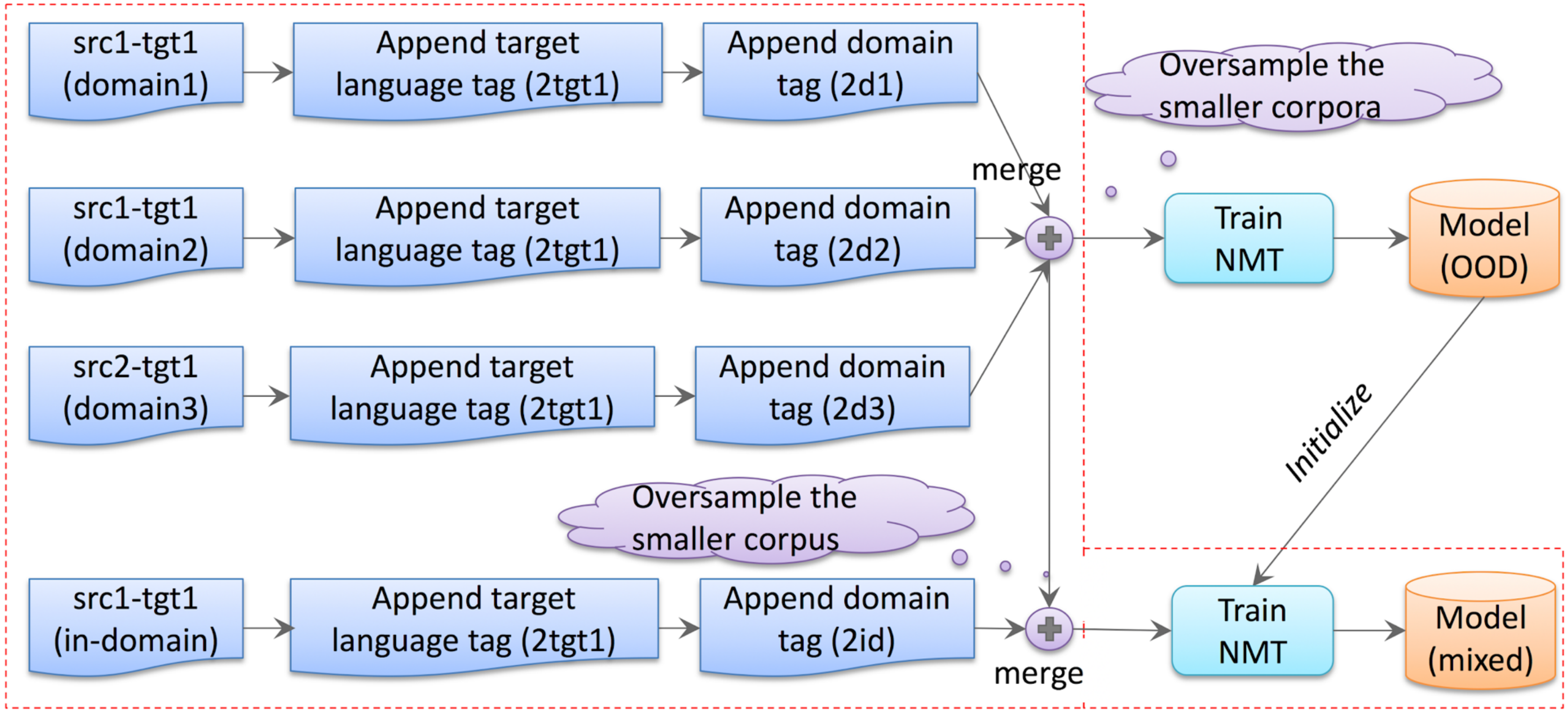
Outline

1. Brief Introduction of Domain Adaptation
2. Domain Adaptation for SMT
3. Domain Adaptation for NMT
    a. Data Centric
    b. Model Centric
        i. Training Objective Centric: Mixed Fine Tuning
        ii. Architecture Centric
        iii. Decoding Centric
4. Domain Adaptation in Specific Scenarios
5. Datasets and Resources
6. Future Directions

Mixed Fine Tuning [Chu+ 2017]



Multilingual and Multi-Domain [Chu+ 2018]



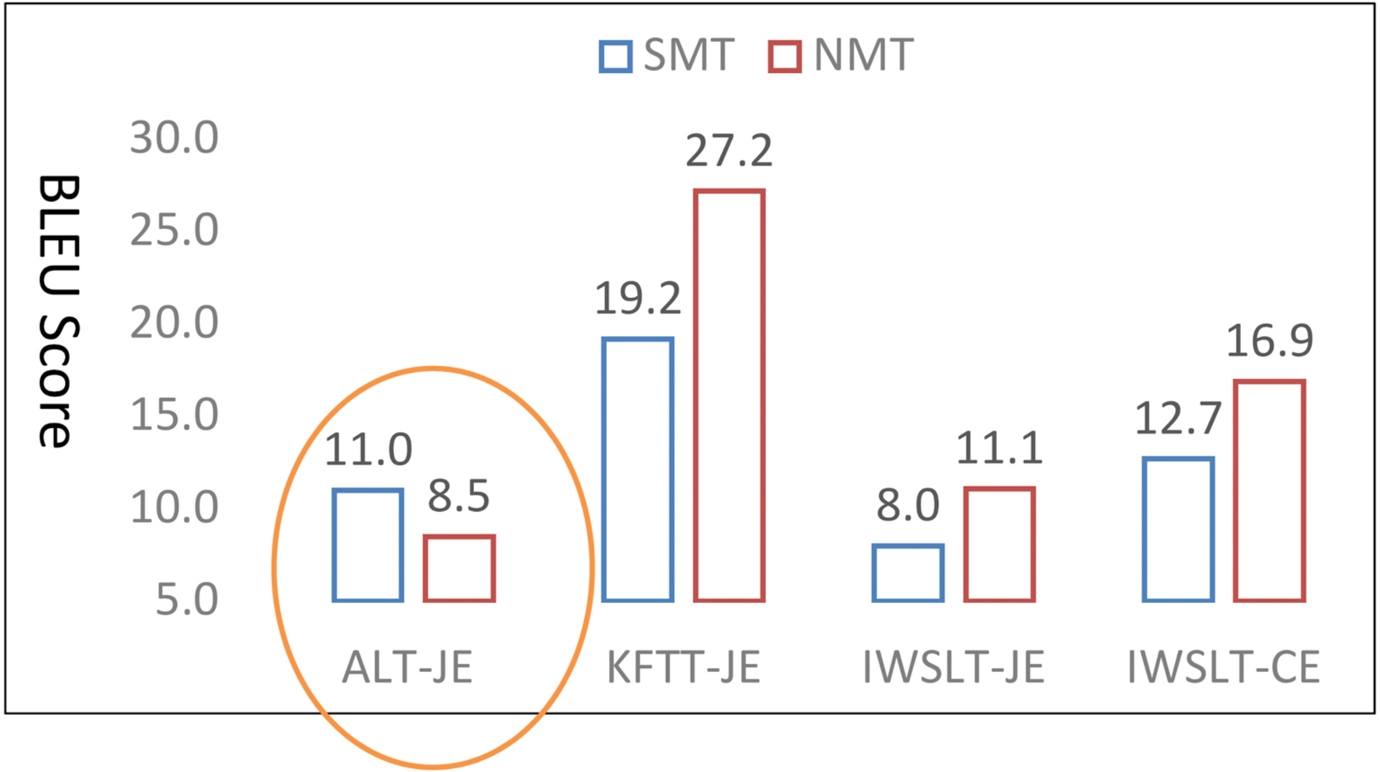
Experimental Settings

- MT tasks

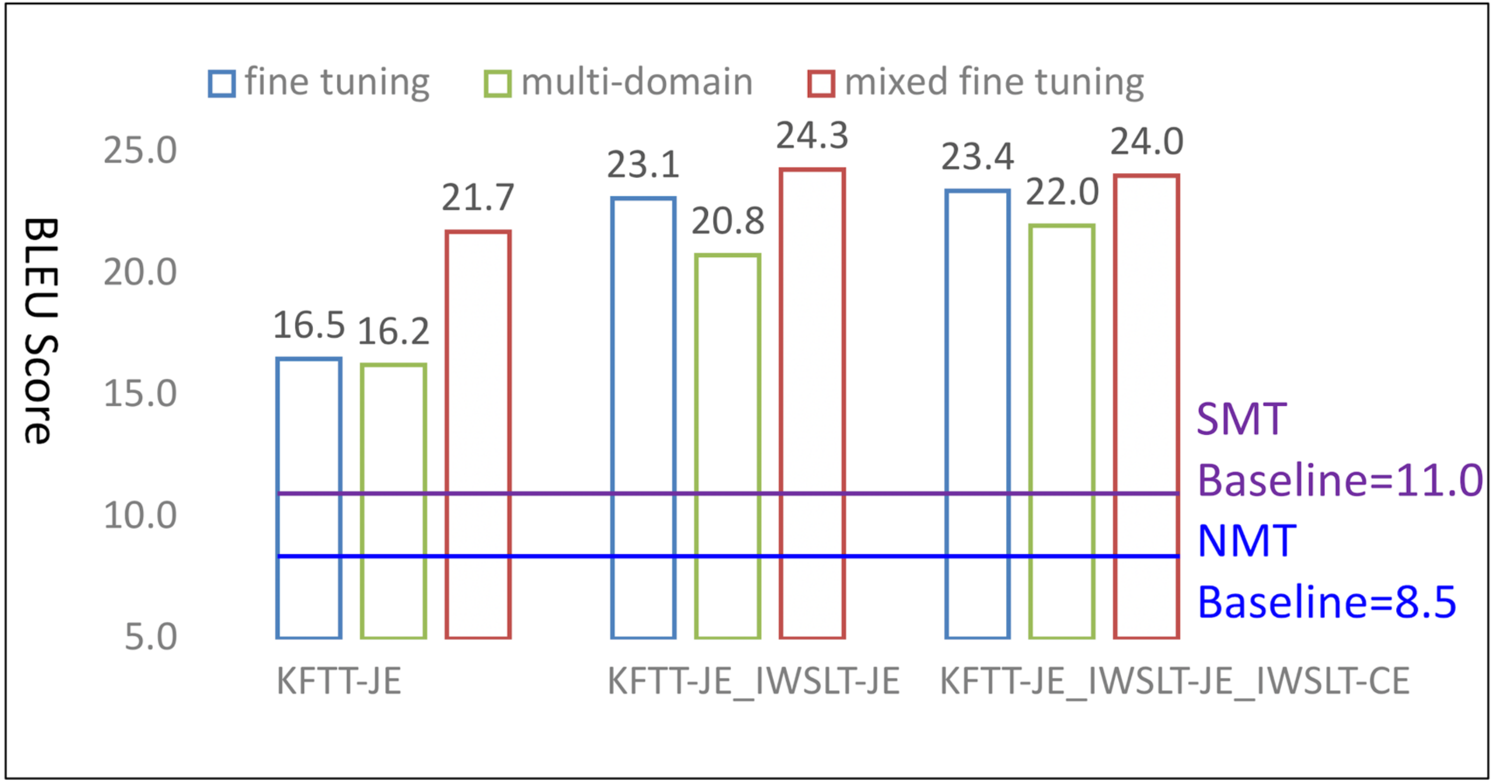
	Corpus (domain)	train	dev	test
In-domain	ALT-JE (Wikinews) [Thu+ 2016]	18k	1,000	1,018
Out-of-domain	KFTT-JE (Wiki-Kyoto) [Neubig+ 2011]	440k	1,166	1,160
	IWSLT-JE (spoken) [Gettolo+ 2015]	223k	871	1,549
	IWSLT-CE (spoken) [Gettolo+ 2015]	209k	887	1,570

- MT systems
  - SMT: Moses [Koehn+ 2007]
  - NMT: Transformer [Vaswani+ 2017]

Results on ALT-JE Without Domain Adaptation



Domain Adaptation Results on ALT-JE



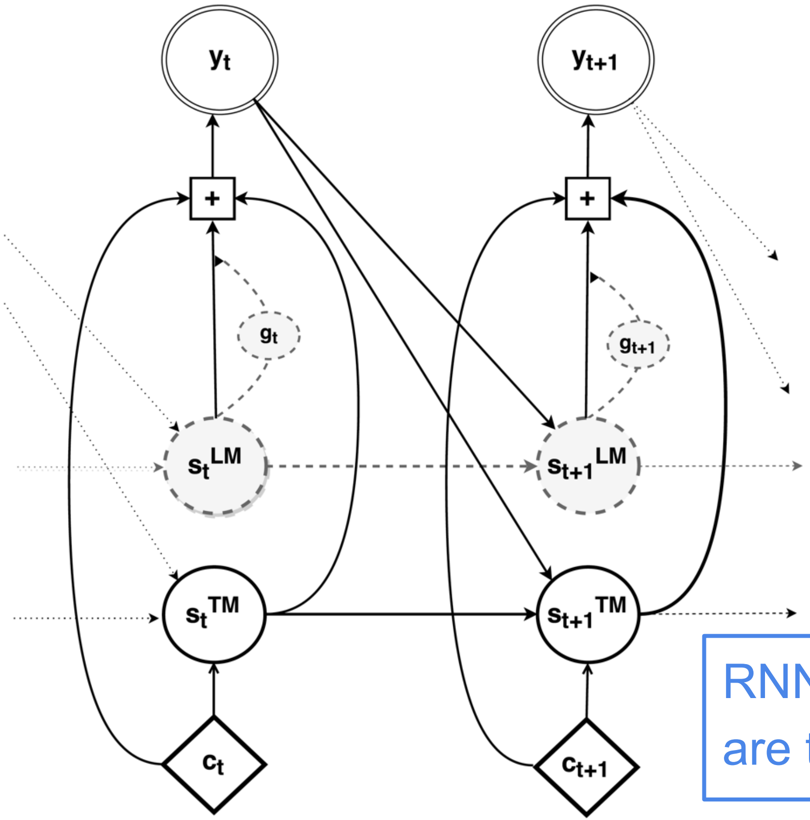
Translation Examples

- Input: シドニーのランドウィック競馬場の８頭のサラブレッド競走馬が馬インフルエンザに感染していることが確認された。
- Reference: it has been confirmed that eight thoroughbred race horses at randwick racecourse in sydney have been infected with equine influenza.
- NMT baseline: the thoroughbred has been confirmed to be infected with the kurawicked when the thoroughbred was infected.
- Fine tuning: it was confirmed that the eight main randwick service predominantly was infected by horse flu.
- Multi-domain: sydney's eight horsthoroughbourghbours were confirmed to be infected with influenza at the horse.
- Mixed fine tuning: it was confirmed that the eight thoroughbred horse racing at the sydney's randowic race course was infected with horse flu.

Outline

1. Brief Introduction of Domain Adaptation
2. Domain Adaptation for SMT
3. Domain Adaptation for NMT
    a. Data Centric
    b. Model Centric
        i. Training Objective Centric
        ii. Architecture Centric: Deep Fusion
        iii. Decoding Centric
4. Domain Adaptation in Specific Scenarios
5. Datasets and Resources
6. Future Directions

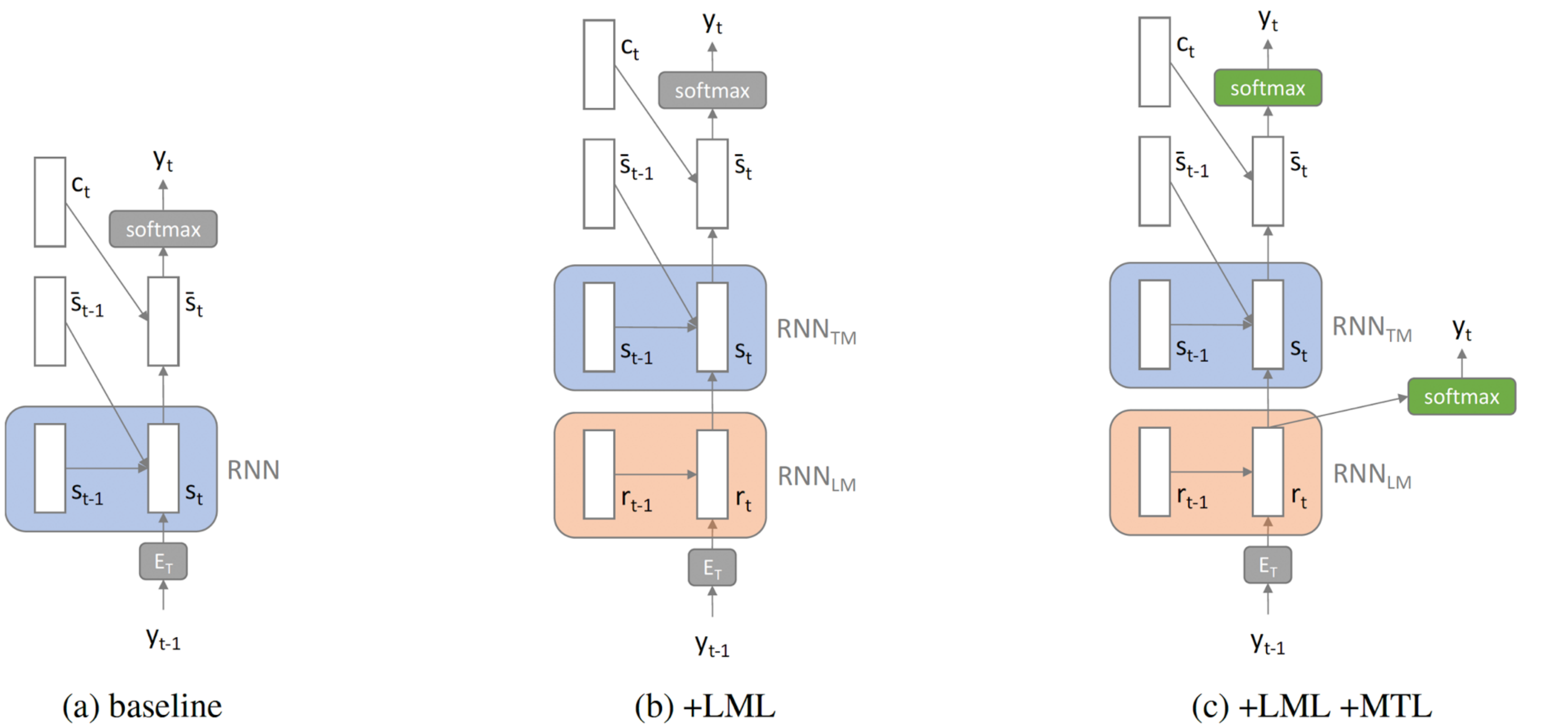
Deep Fusion (1/2) [Gulcehre+ 2015]



RNNLM and NMT models are trained separately

Deep Fusion (2/2) [Domhan+ 2017]

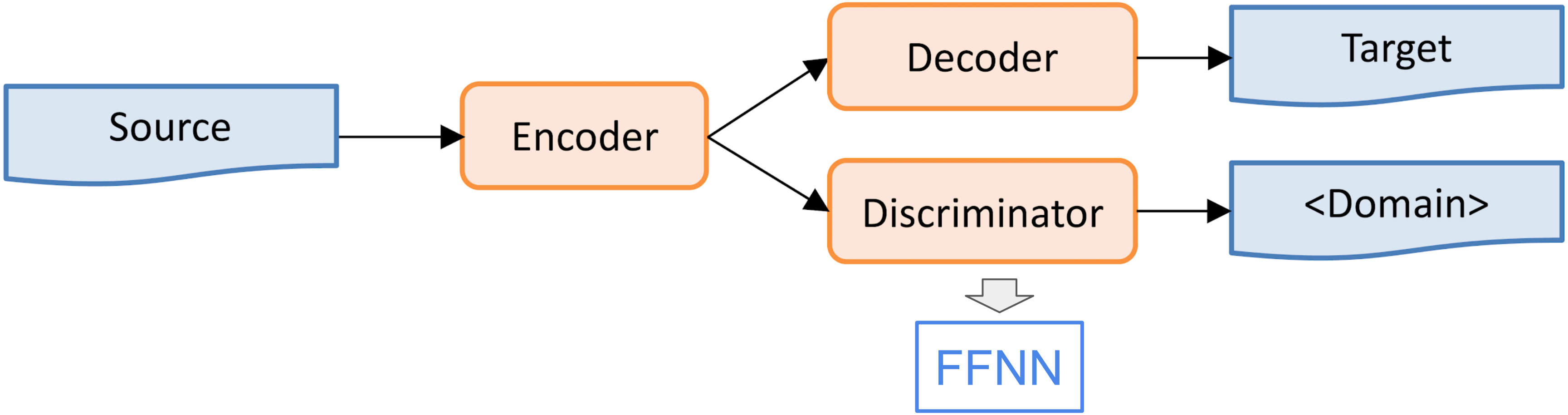
RNNLM and NMT models are trained jointly



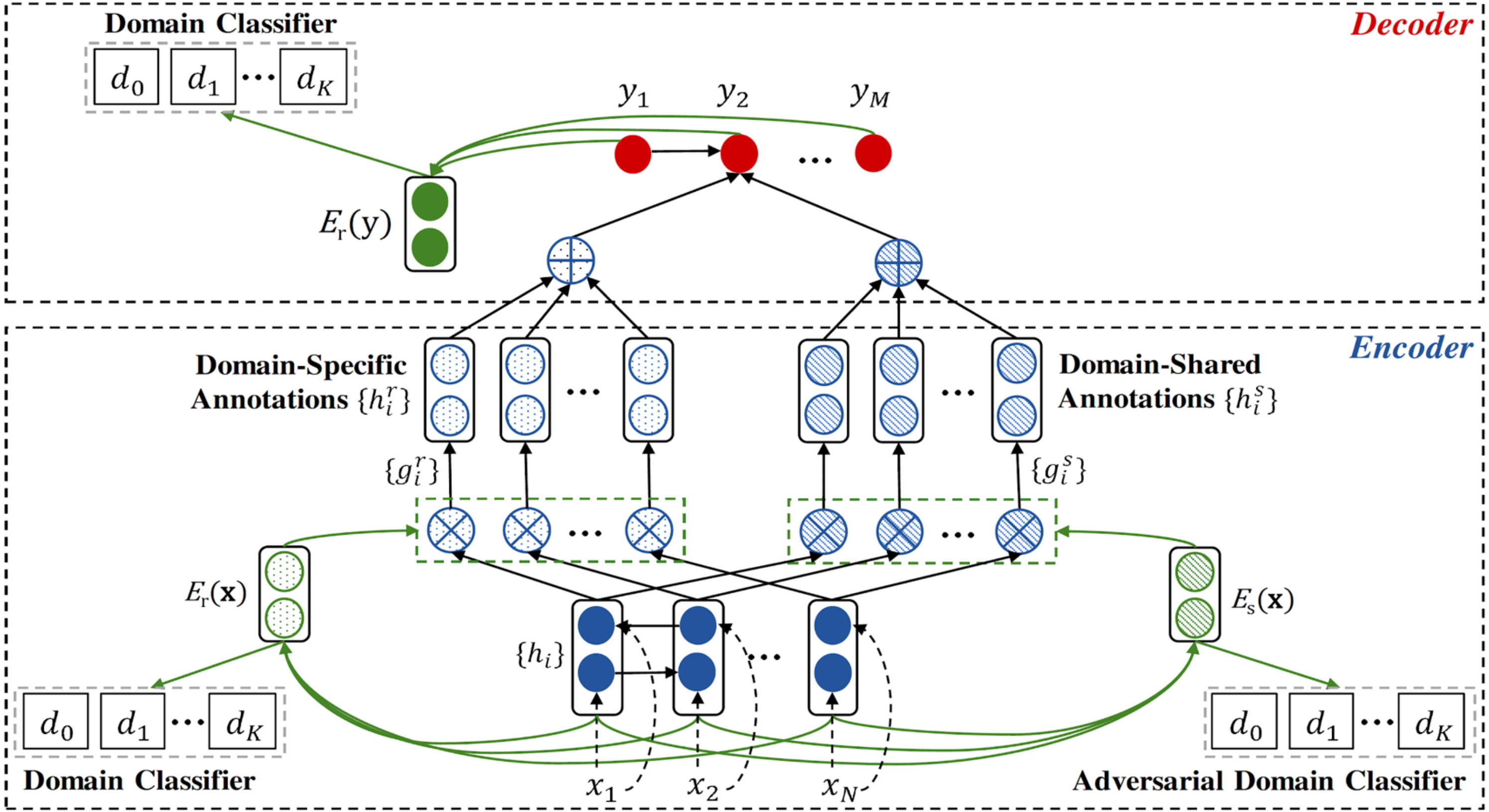
Outline

1. Brief Introduction of Domain Adaptation
2. Domain Adaptation for SMT
3. Domain Adaptation for NMT
    a. Data Centric
    b. Model Centric
        i. Training Objective Centric
        ii. Architecture Centric: Domain Discriminator
        iii. Decoding Centric
4. Domain Adaptation in Specific Scenarios
5. Datasets and Resources
6. Future Directions

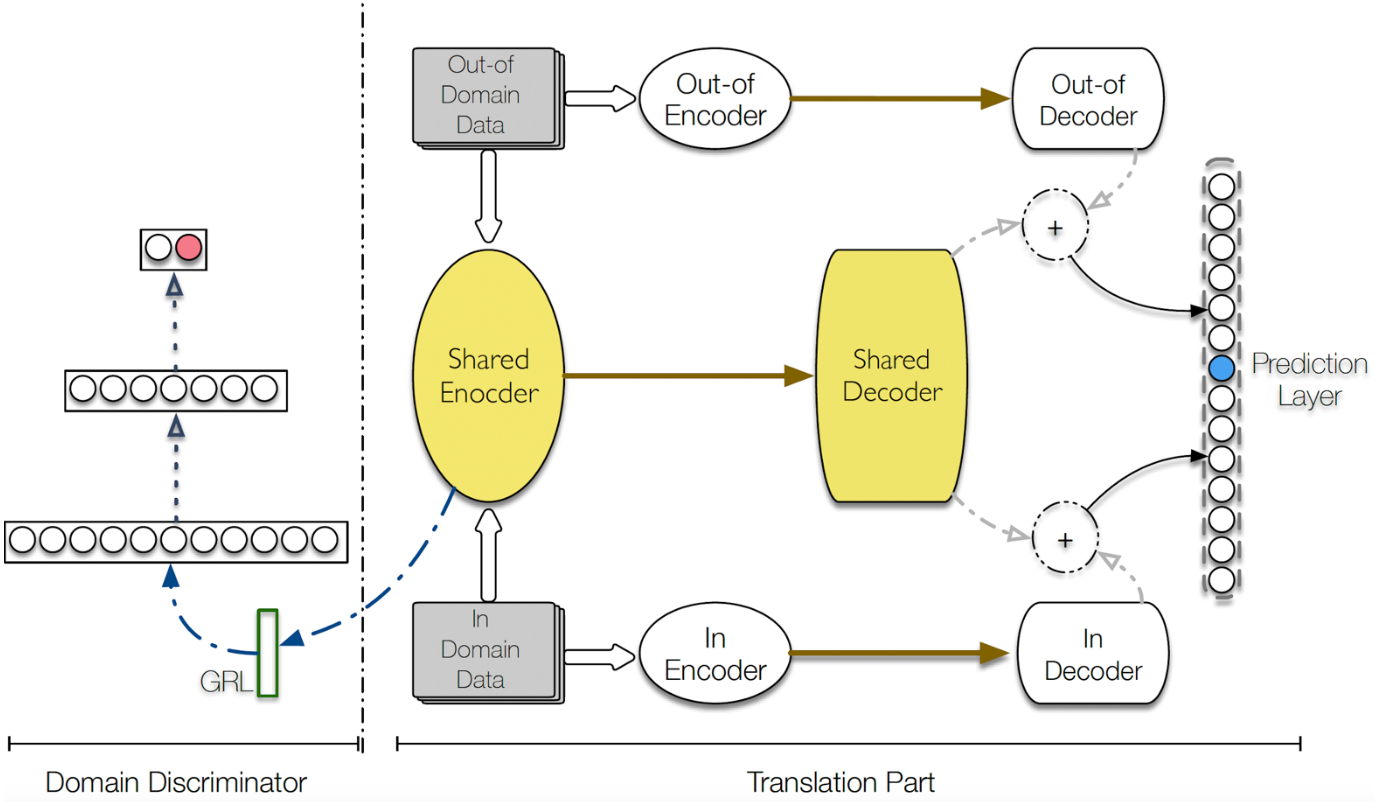
Domain Discriminator [Britz+ 2017]



Word-Level Domain Discriminator [Zeng+ 2018]



Multiple Encoders and Decoders [Gu+ 2019]



Results of Multiple Encoders and Decoders [Gu+ 2019]

En-Zh	dev	test	average
In	32.45	30.42	31.44
Out + In	30.37	28.76	29.57
Multi-domain: Sampler	35.06	32.97	34.02
Fine Tune	35.02	33.36	34.19
Tag: DC	31.08	29.59	30.34
[Britz+ 2017]: DM	30.98	29.73	30.36
TTM	31.77	30.11	30.94
ADM	31.23	29.88	30.56
our method	36.55**	34.84**	35.70

Table 1: Results of the en-zh translation experiments.

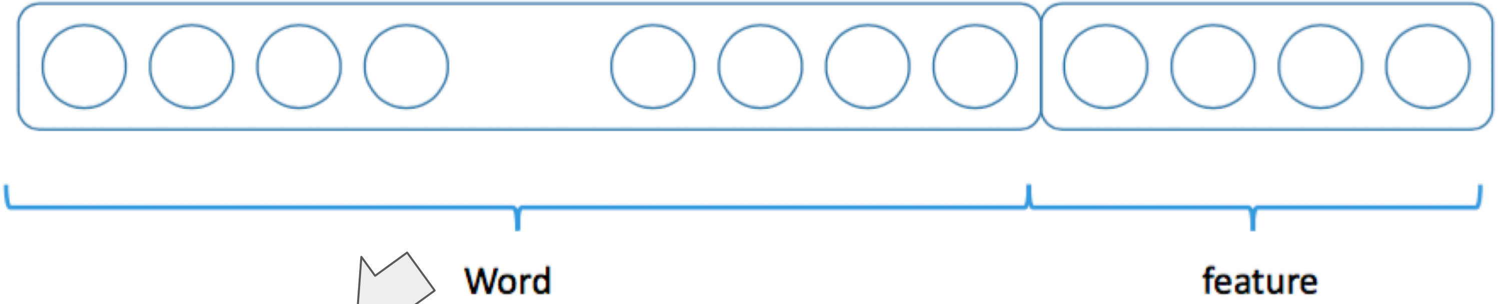
En-De	test06	test07	average
In	23.36	25.00	24.18
Out + In	20.69	22.43	21.56
Sampler	26.83	29.01	27.92
Fine Tune	27.02	29.19	28.11
our method	27.97*	30.67**	29.32

Table 2: Results of the WMT 07 en-de translation experiments.

Outline

1. Brief Introduction of Domain Adaptation
2. Domain Adaptation for SMT
3. Domain Adaptation for NMT
    a. Data Centric
    b. Model Centric
        i. Training Objective Centric
        ii. Architecture Centric: Domain Control
        iii. Decoding Centric
4. Domain Adaptation in Specific Scenarios
5. Datasets and Resources
6. Future Directions

Domain Control [Kobus+ 2016]



Src: `Headache|MED may|MED be|MED experienced|MED`

Trg: `Des céphalées peuvent survenir`

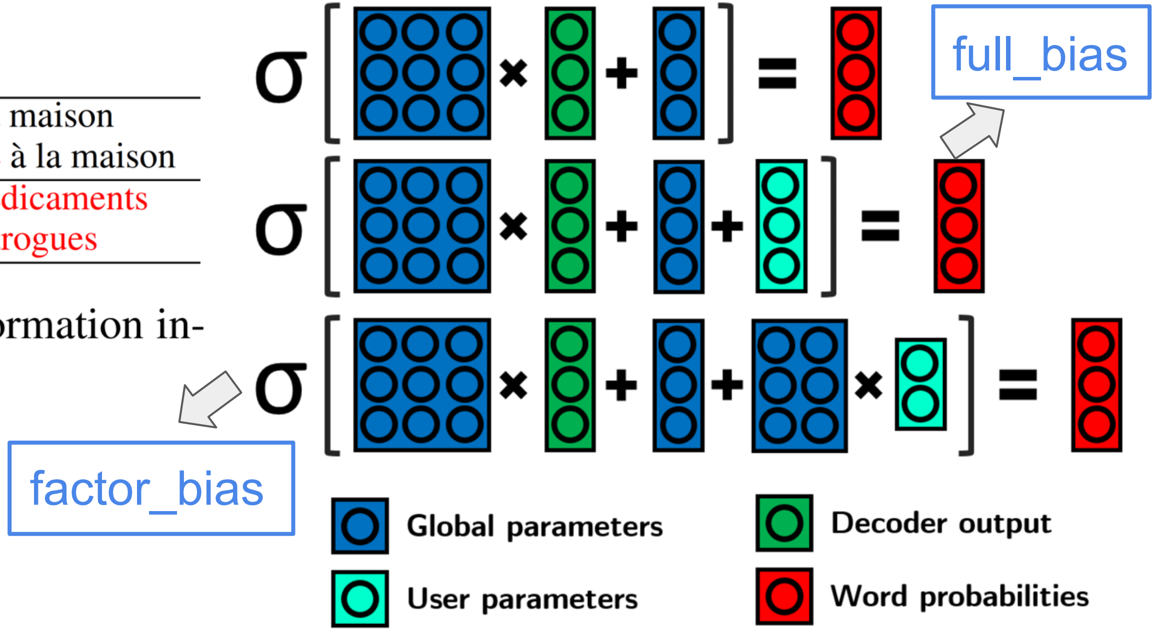
Src: `Headache may be experienced` **`@MED@`**

Tgt: `Des céphalées peuvent survenir`

Extreme Adaptation [Michel+ 2018]

Source	Translation
I went home	[Man]: Je suis rentré à la maison [Woman]: Je suis rentrée à la maison
I do drug testing	[Doctor]: Je teste des médicaments [Police]: Je dépiste des drogues

Table 1: Examples where speaker information influences English-French translation.



Results of Extreme Adaptation [Michel+ 2018]

	en-fr	en-es	en-de
base	38.05	39.89	26.46
spk_token	**38.85**	40.04	26.52
full_bias	**38.54**	**40.30**	**27.20**
fact_bias	**39.01**	39.88	**26.94**

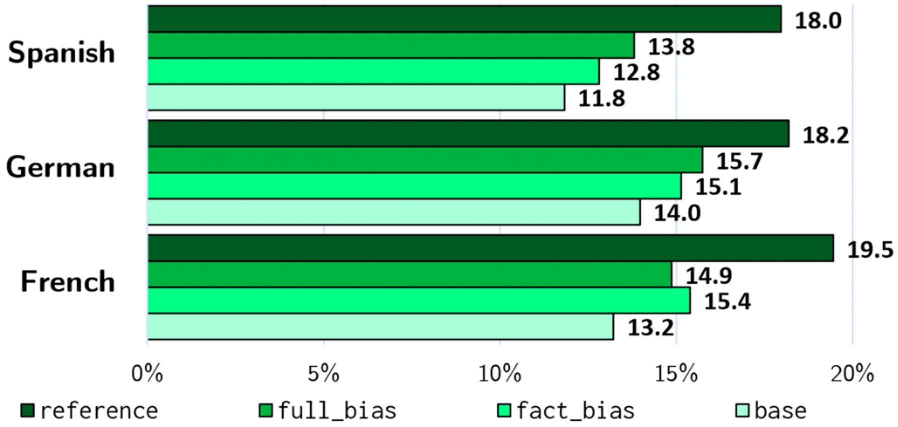
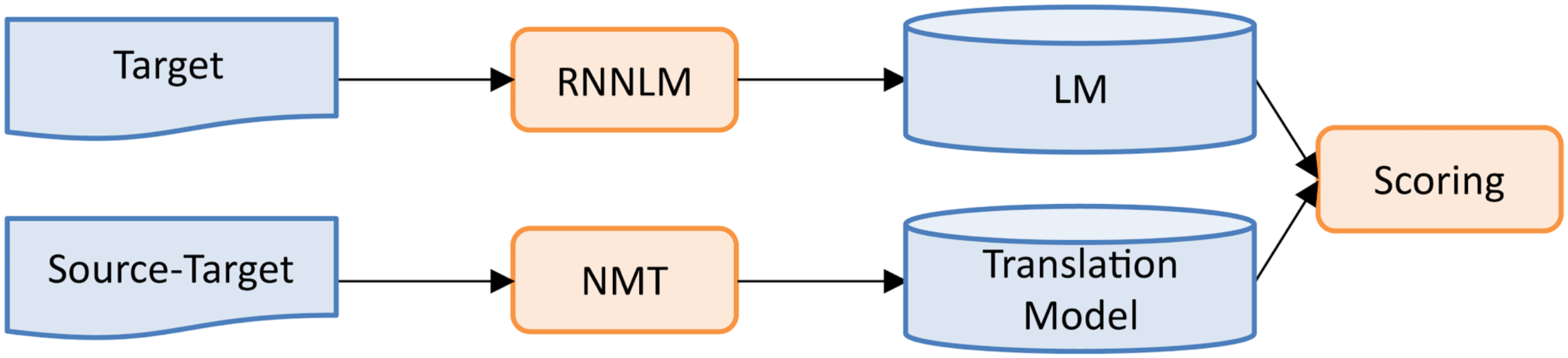


Figure 2: Speaker classification accuracy of our continuous bag-of-n-grams model.

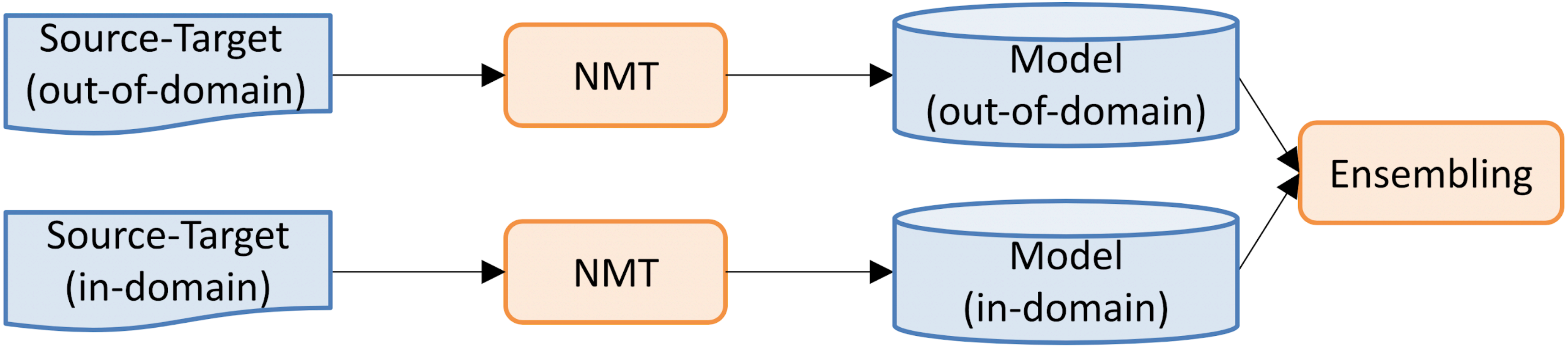
Outline

1. Brief Introduction of Domain Adaptation
2. Domain Adaptation for SMT
3. Domain Adaptation for NMT
    a. Data Centric
    b. Model Centric
        i. Training Objective Centric
        ii. Architecture Centric
        iii. Decoding Centric
4. Domain Adaptation in Specific Scenarios
5. Datasets and Resources
6. Future Directions

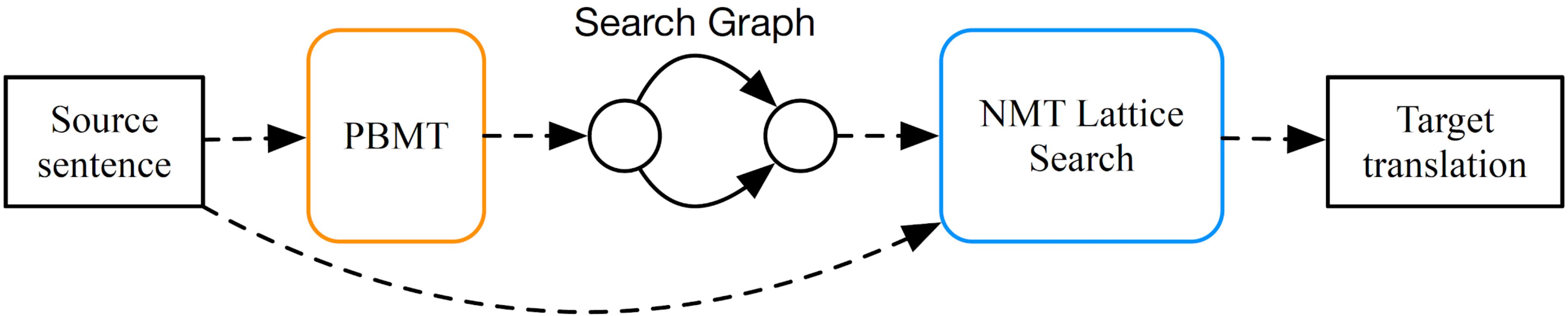
Shallow Fusion [Gulcehre+ 2015]



Ensembling [Freitag+ 2016]



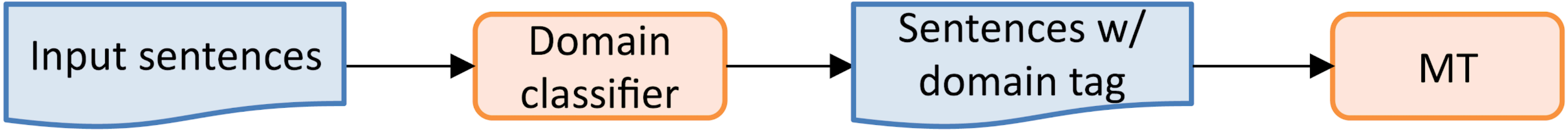
Neural Lattice Search [Khayrallah+ 2017]



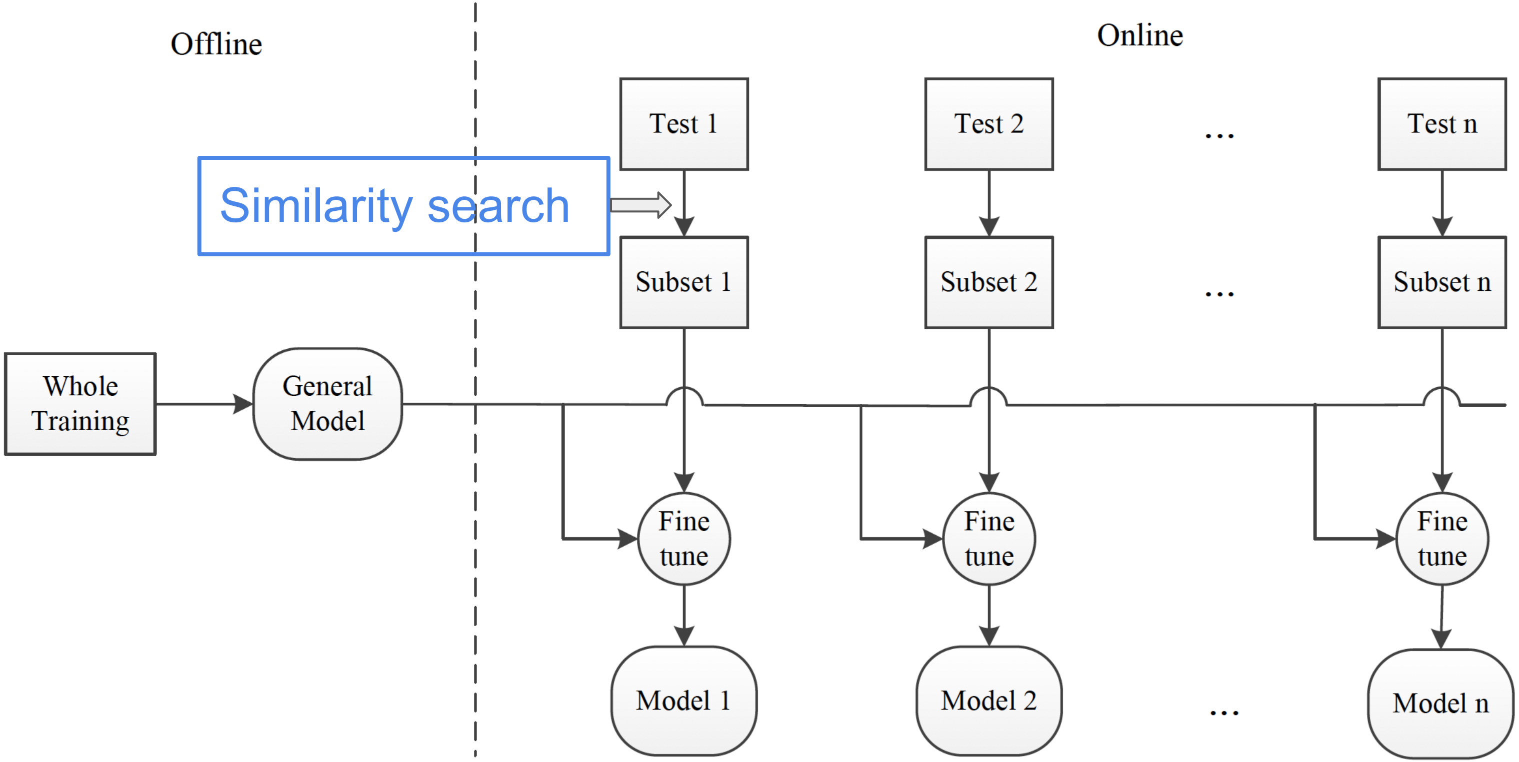
Outline

1. Brief Introduction of Domain Adaptation
2. Domain Adaptation for SMT
3. Domain Adaptation for NMT
4. Domain Adaptation in Specific Scenarios
    a. Input Domain Unknown
    b. Incremental/Online Domain Adaptation
5. Datasets and Resources
6. Future Directions

Input Domain Unknown



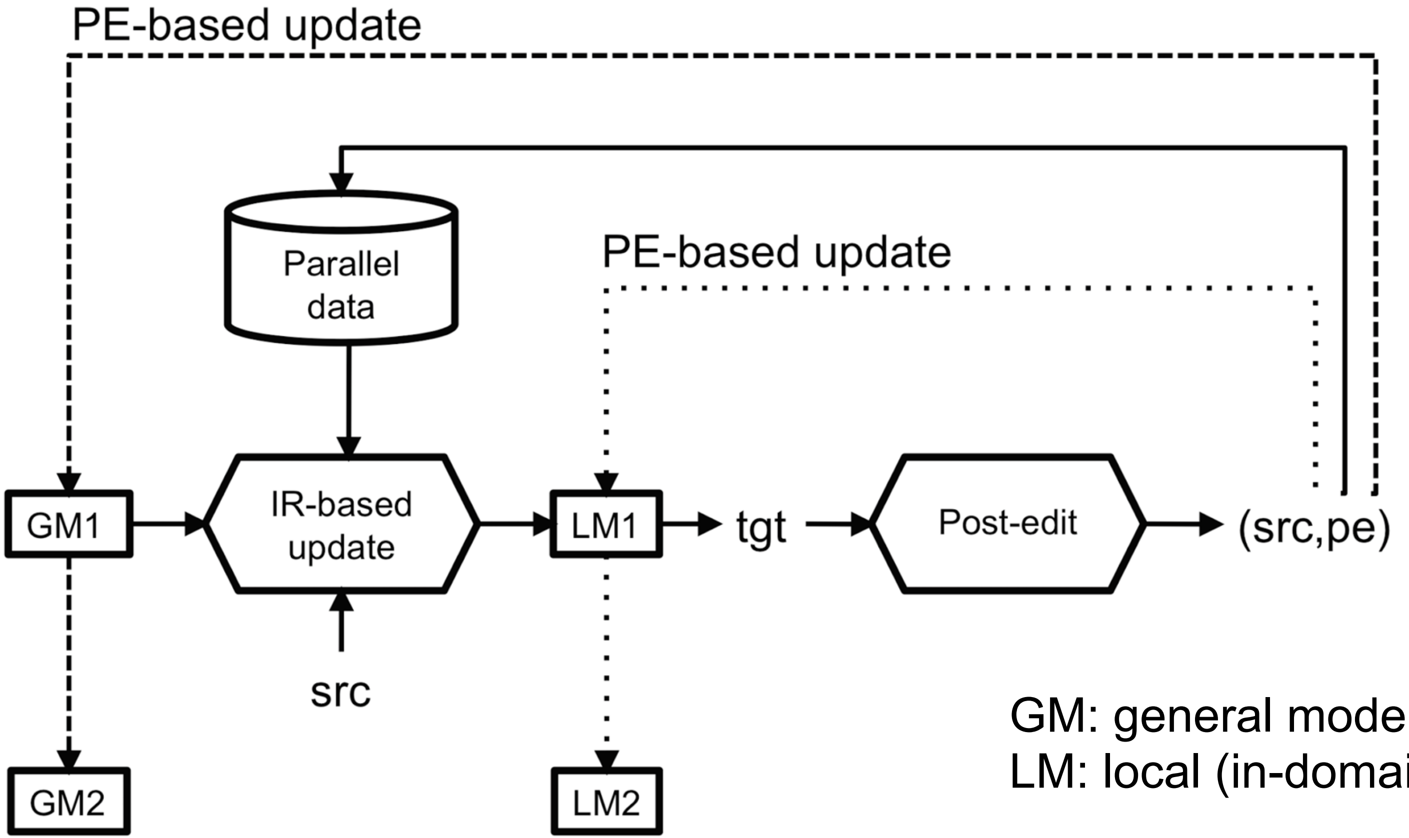
Sentence Retrieve Based Model [Li+ 2016]



Outline

1. Brief Introduction of Domain Adaptation
2. Domain Adaptation for SMT
3. Domain Adaptation for NMT
4. Domain Adaptation in Specific Scenarios
    a. Input Domain Unknown
    b. Incremental/Online Domain Adaptation
5. Datasets and Resources
6. Future Directions

Post-edit for Online Domain Adaptation [Turchi+ 2017]



GM: general model
LM: local (in-domain) model

Outline

1. Brief Introduction of Domain Adaptation
2. Domain Adaptation for SMT
3. Domain Adaptation for NMT
4. Domain Adaptation in Specific Scenarios
5. Datasets and Resources
6. Future Directions

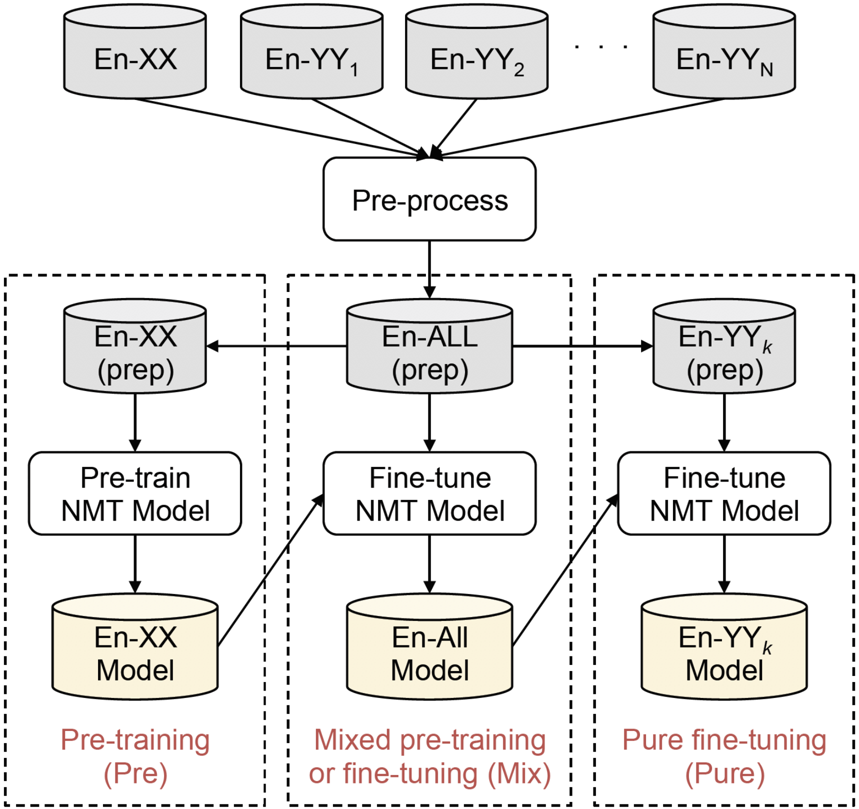
Datasets and Resources

Studies	Language pairs	In-domain corpora	Out-of-domain corpora
Wang et al. [114, 115, 116]	En-De & En-Fr	IWSLT	WMT
Chu et al. [19]	Zh-En	IWSLT	NTCIR
	Zh-Ja	Wiki-CJ	ASPEC
Chen et al. [11]	Zh-En	Dev data of NIST	Training data of NIST
	En-Fr	Dev data of WMT	Training data of WMT
Michel and Neubig [79]	En-Fr & En-De & En-Es	Certain speaker of IWSLT	IWSLT
van der Wees et al. [110]	En-De	Dev data of TED &WMT & Movie dialogues & EMEA medical	Training data of TED &WMT & Movie dialogues & EMEA medical
Zhang et al. [130]	En-De & Ru-En	IWSLT & Patents	Paracrawl
Farajian et al. [33]	En-Fr	Multi-domain	Multi-domain
Zhang et al. [130]	Ru-En & En-De	IWSLT & Patent	Web-crawled
Gu et al. [42]	En-Zh	Laws	LDC
	En-De	News Commentary	WMT (NEWS)

Outline

1. Brief Introduction of Domain Adaptation
2. Domain Adaptation for SMT
3. Domain Adaptation for NMT
4. Domain Adaptation in Specific Scenarios
5. Datasets and Resources
6. Future Directions
    a. Multilingual and Multi-Domain Adaptation
    b. Unsupervised NMT

Multi-stage Learning [Dabre+ 2019]



Results of Multi-stage Learning [Dabre+ 2019]

#	XX	N	Model capacity	Training configuration			YY test set						
				Pre	Mix	Pure	Bn	Tl	Id	Ja	Km	Ms	Vi
1.	-	1	1-to-1	-	-	✓	3.99	24.04	24.10	11.03	22.53	29.85	27.39
2.	Zh	1	1-to-2	✓	-	✓	8.86*	27.54*	27.10*	19.07*	28.41*	32.52*	34.63*
3.	Zh	1	1-to-2	-	✓	✓	4.90*	23.07	23.37	13.97*	26.13*	29.24	29.82*
4.	Zh	1	1-to-2	✓	✓	✓	7.99*	26.61*	25.62*	18.39*	27.49*	31.63*	34.22*
5.	Zh	7	1-to-8	✓	-	✓	8.54*	26.88*	26.02*	18.99*	27.07*	32.39*	33.32*
6.	Zh	7	1-to-8	-	✓	✓	9.43*	25.86*	26.33*	19.34*	26.86*	32.39*	33.28*
7.	Zh	7	1-to-8	✓	✓	✓	**10.30**$^{*+\dagger}$	**28.22**$^{*+\dagger}$	**27.24***†	**20.08**$^{*+\dagger}$	**28.66***†	**33.19**$^{*+\dagger}$	**35.34**$^{*+\dagger}$
2.	Ja	1	1-to-2	✓	-	✓	9.16*	28.06*	26.53*	21.55*	27.98*	33.68*	33.93*
3.	Ja	1	1-to-2	-	✓	✓	4.37	22.91	23.37	16.47$*$	23.36$*$	29.28	29.10$*$
4.	Ja	1	1-to-2	✓	✓	✓	8.77*	26.64*	25.88*	21.61*	27.55*	32.45*	34.29*
5.	Ja	7	1-to-8	✓	-	✓	9.43*	27.45*	26.70*	21.79*	27.87*	32.92*	34.28*
6.	Ja	7	1-to-8	-	✓	✓	9.96^{*+}	28.39*	27.22^{*+}	21.03*	28.91^{*+}	33.75*	36.00^{*+}
7.	Ja	7	1-to-8	✓	✓	✓	**10.77**$^{*+\dagger}$	**28.62**$^{*+}$	**28.89**$^{*+\dagger}$	**22.60**$^{*+\dagger}$	**30.03**$^{*+\dagger}$	**34.75**$^{*+\dagger}$	**37.06**$^{*+\dagger}$

Outline

1. Brief Introduction of Domain Adaptation
2. Domain Adaptation for SMT
3. Domain Adaptation for NMT
4. Domain Adaptation in Specific Scenarios
5. Datasets and Resources
6. Future Directions
    a. Multilingual and Multi-Domain Adaptation
    b. Unsupervised NMT

Domain Adaptation for Unsupervised NMT

Scenarios for unsupervised NMT ate different from supervised NMT

Scenarios	Abbreviation	L_1 in-domain	L_2 in-domain	L_1 out-of-domain	L_2 out-of-domain
Monolingual corpora from same domains	II	✓	✓	×	×
	OO	×	×	✓	✓
	$IIOO$	✓	✓	✓	✓
Monolingual corpora from different domains	IOO	×	✓	✓	✓
	IIO	✓	✓	✓	×
	IO	×	✓	✓	×

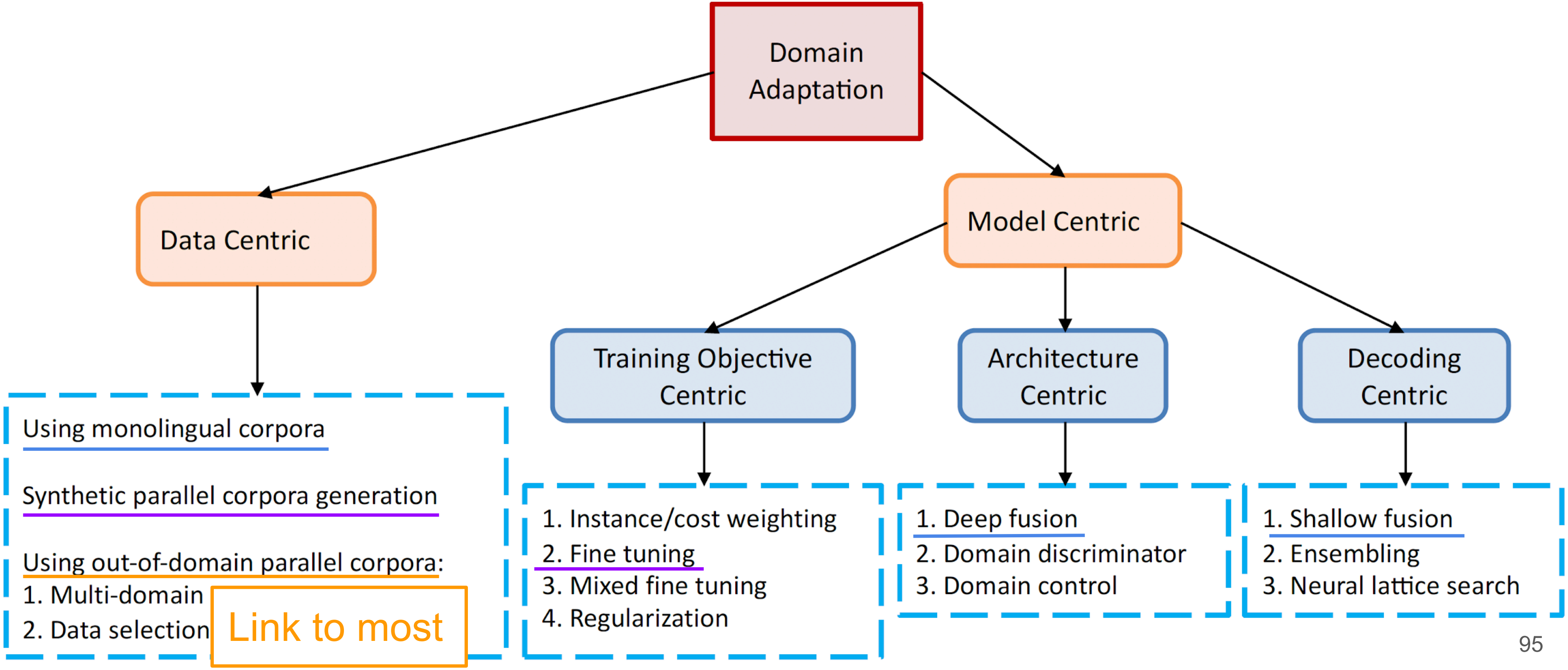
Some Initial Results [Sun+ 2019]

Scenario	Supervision	Method	De-En		En-De		Fr-En		En-Fr		#
			test2012	test2013	test2012	test2013	test2010	test2011	test2010	test2011	
II	Yes	Wang et al. (2018)	n/a	n/a	23.07	25.40	n/a	n/a	32.11	35.22	1
		Base	33.68	35.41	28.09	30.48	36.13	40.07	36.43	37.58	2
II	No	Base	24.42	25.65	21.99	22.72	25.94	29.73	25.32	27.06	3
OO		Base	21.21	21.66	10.25	9.90	24.28	28.77	23.08	26.08	4
IIOO	No								26.35	30.12	5
									29.08	33.67	6
IOO	No								25.18	28.73	7
		FT	22.75	23.14	21.09	21.78	28.37	33.57	26.16	30.14	8
IIO	No	Base	11.11	10.30	11.54	11.95	17.88	20.32	17.02	18.16	9
		FT+BW	26.12	27.33	22.63	23.72	27.88	32.16	25.42	28.05	10
IO	No	Base	10.79	10.77	11.44	11.82	18.00	20.91	16.19	16.84	11
		BW	17.78	18.00	16.01	16.60	22.53	25.29	20.04	22.12	12

Existing domain adaptation methods still work but perform differently in different scenarios

BW: batch weighting

Conclusion



Domain Adaptation
Data Centric
Model Centric
Training Objective Centric
Architecture Centric
Decoding Centric
Using monolingual corpora
Synthetic parallel corpora generation
Using out-of-domain parallel corpora:
1. Multi-domain
2. Data selection
Link to most
1. Instance/cost weighting
2. Fine tuning
3. Mixed fine tuning
4. Regularization
1. Deep fusion
2. Domain discriminator
3. Domain control
1. Shallow fusion
2. Ensembling
3. Neural lattice search