Domain Adaptation for Machine Translation

Mieradilijiang Maimaiti

2017-10-26

Outline

- Introduction
  - Domain adaptation
  - Machine translation
- Domain Adaptation for SMT
  - Self-training
  - Data selection
  - Data weighting
  - Context based
  - Topic based
- Domain Adaptation for NMT
- Our work
- Conclusion && Future work

Outline

- Introduction
  - Domain adaptation
  - Machine translation
- Domain Adaptation for SMT
  - Self-training
  - Data selection
  - Data weighting
  - Context based
  - Topic based
- Domain Adaptation for NMT
- Our work
- Conclusion && Future work

- Domain adaptation
- Machine translation

Domain adaptation

- Not a well defined notion.
- Should be based on some concept of textual similarity
  - Lexical choice
  - Grammar
  - Topic
  - Style
  - Genre
  - Register
  - Intent

Domain adaptation

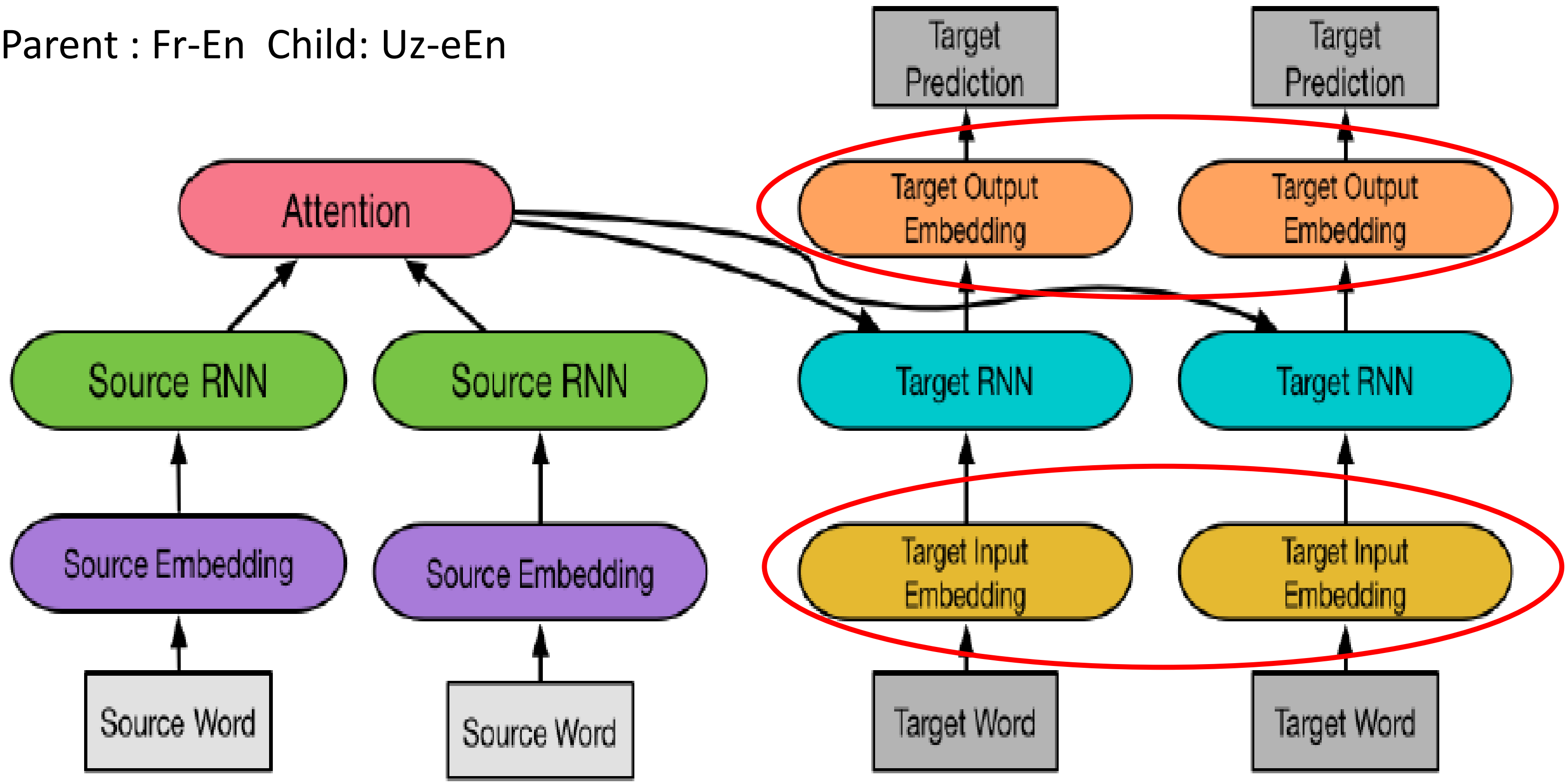
Domain Adaptation (DA) is a field associated with machine learning and transfer learning.

DA is one of the branches of transfer learning.

DA build a system on **one kind of data** and **adjust** it to apply to another.

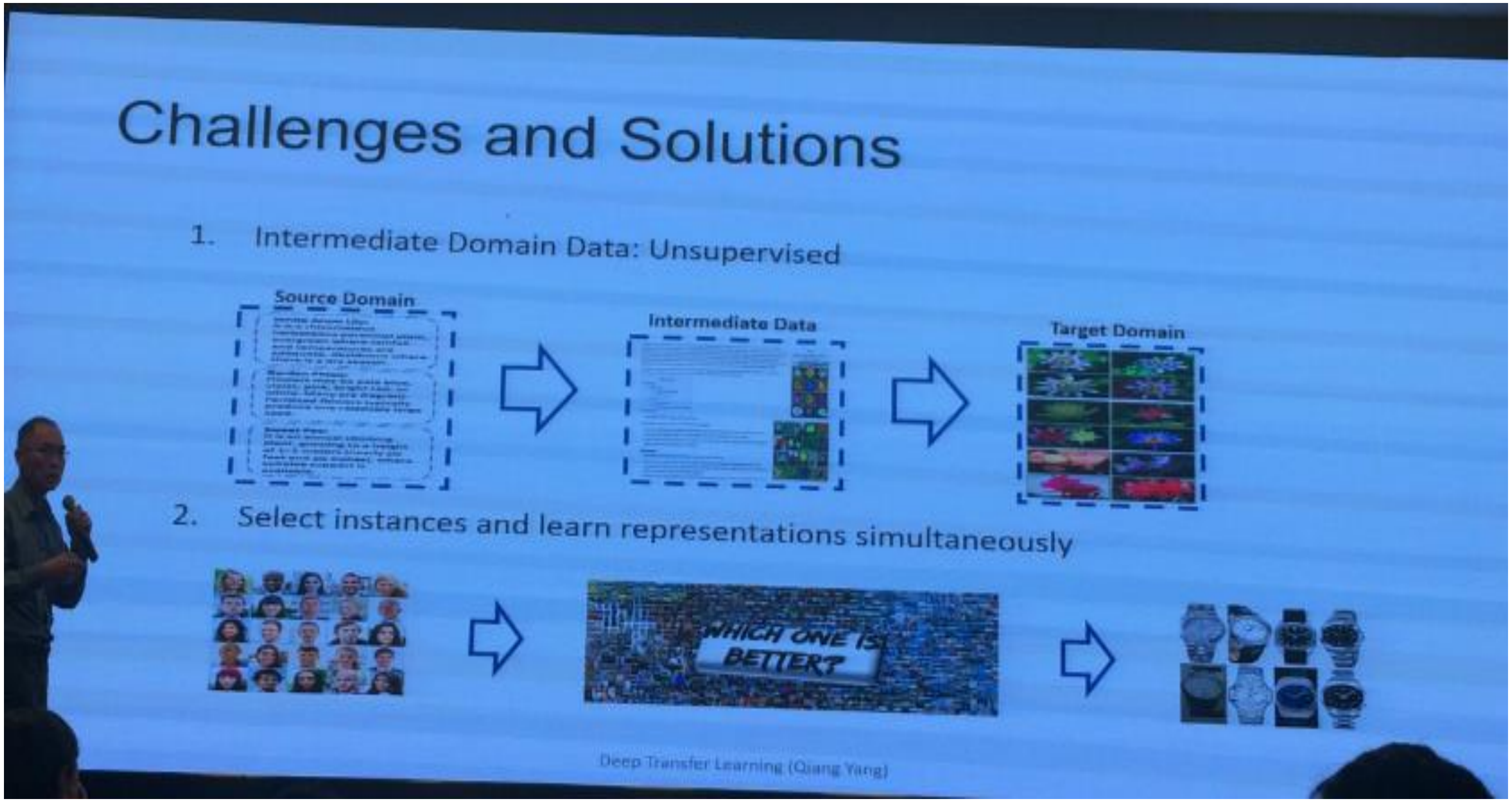
Optimal setting for transferring from **parent** model to **child** model.

Parent : Fr-En Child: Uz-eEn

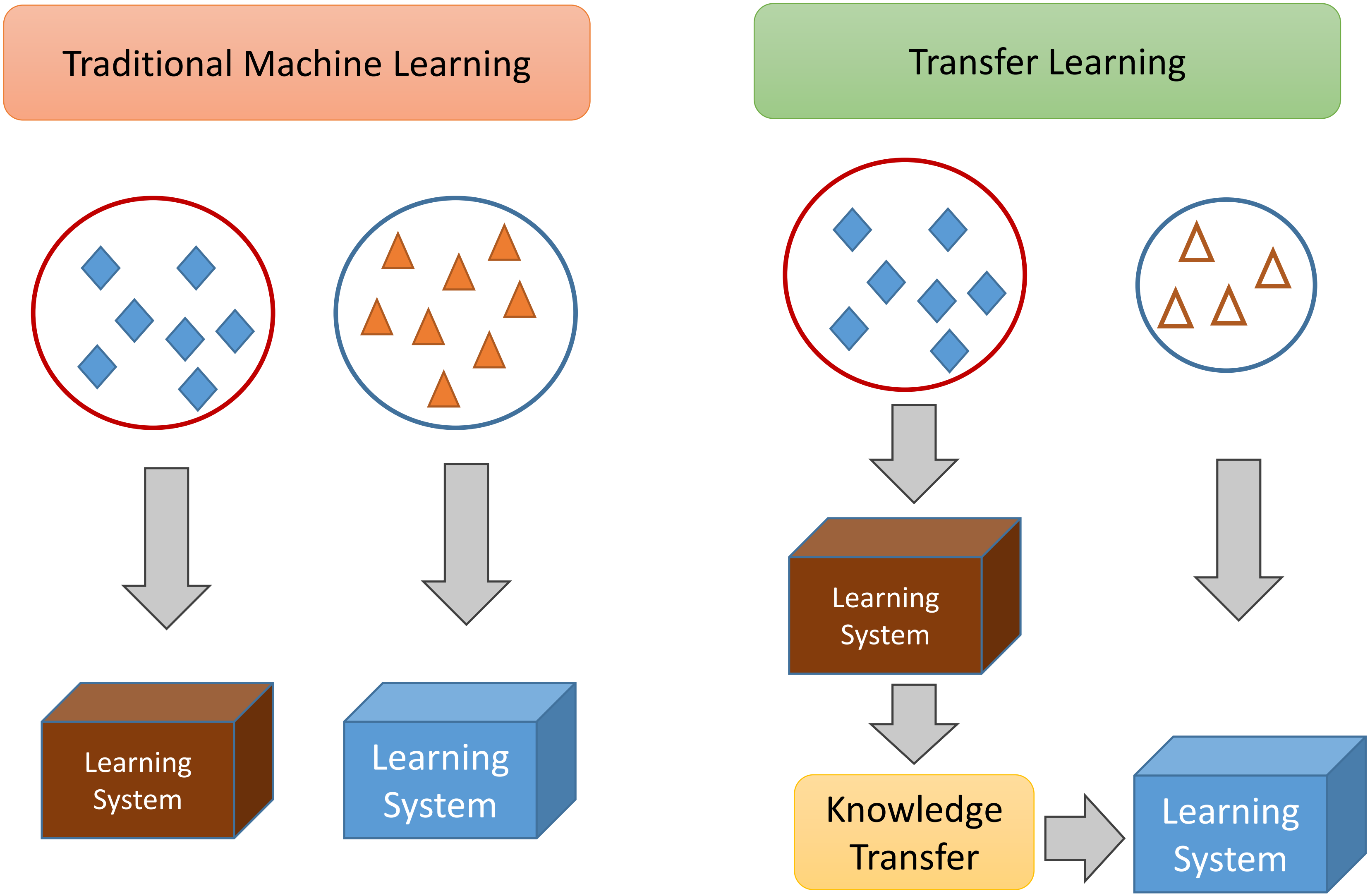


[Barret Zoph et al., 2016]

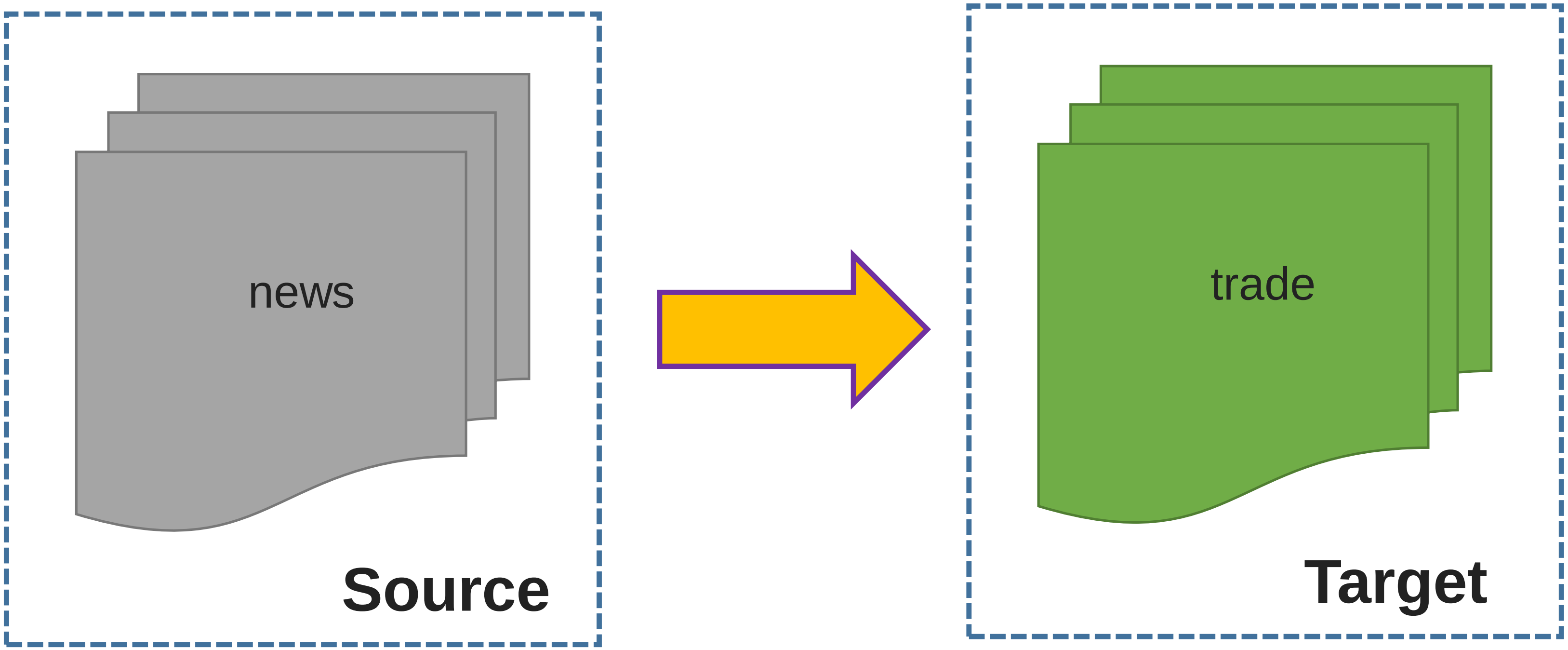
Transfer learning



[Qiang Yang, 2017]


Traditional Machine Learning
Transfer Learning
Learning System
Learning System
Learning System
Knowledge Transfer
Learning System

This scenario arises when we aim at learning from a source data distribution a well performing model on a different (but related) target data distribution.



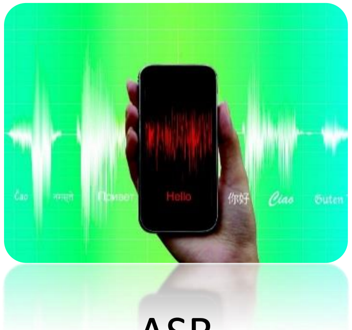
In Natural Language Processing (NLP), train a system on some language data, retune && apply it to specific different task.

Build speech recognition system using recorded phone calls, then tune it to use as an airline reservation hotline.


CV


ER

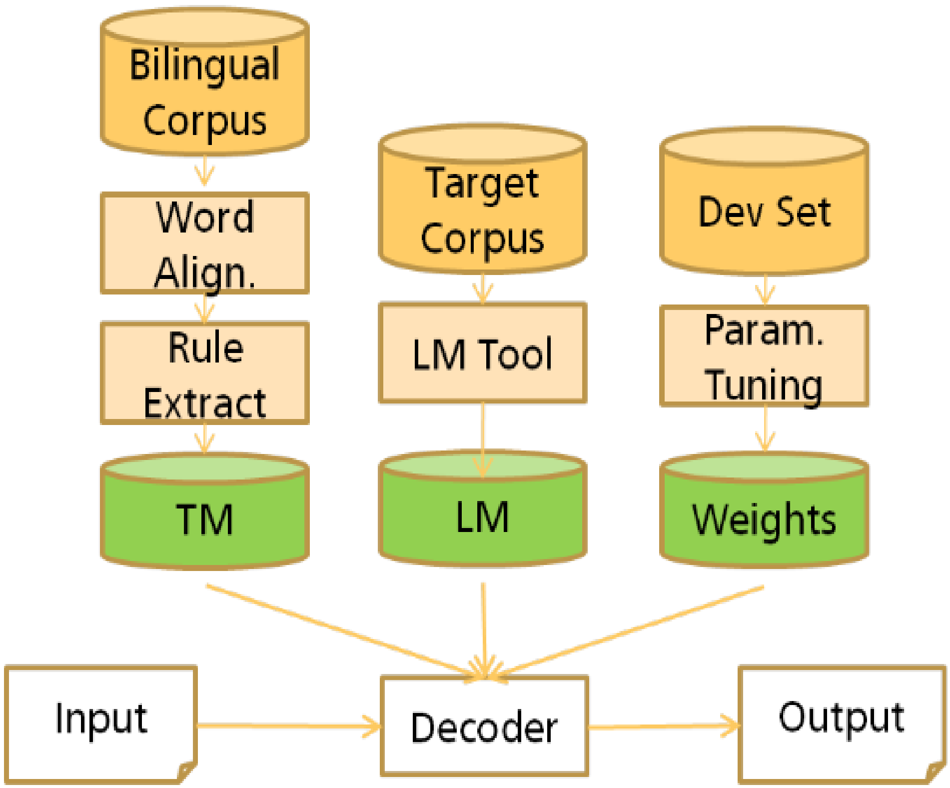
IR


ASR

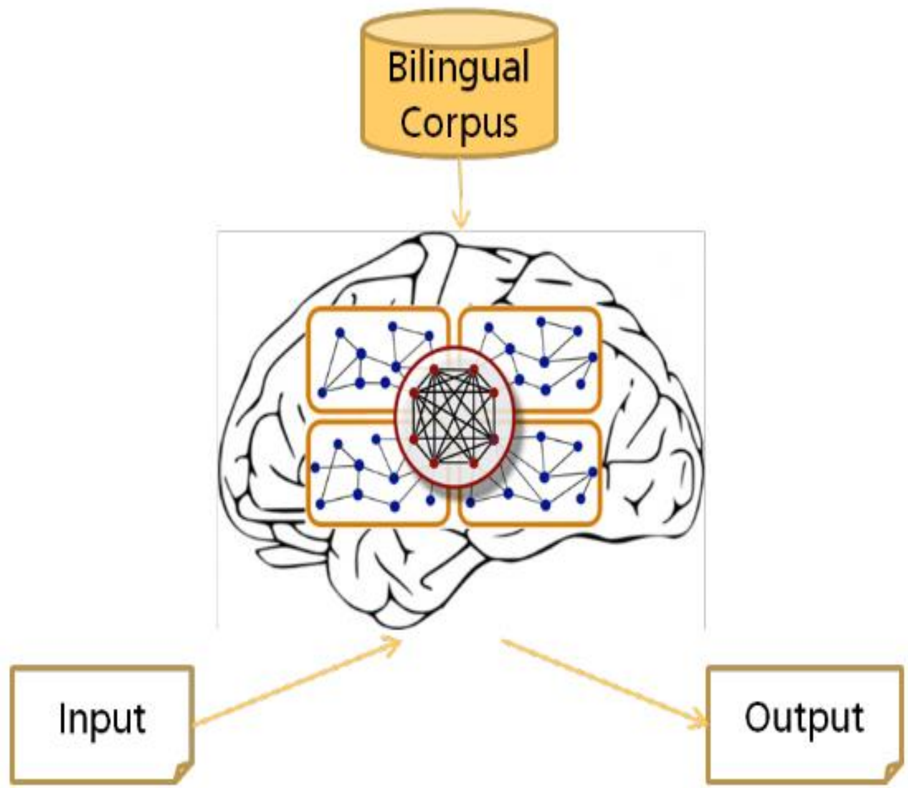
Machine translation

Many sub-components are tuned separately

single , large neural network



SMT (1993 ~)

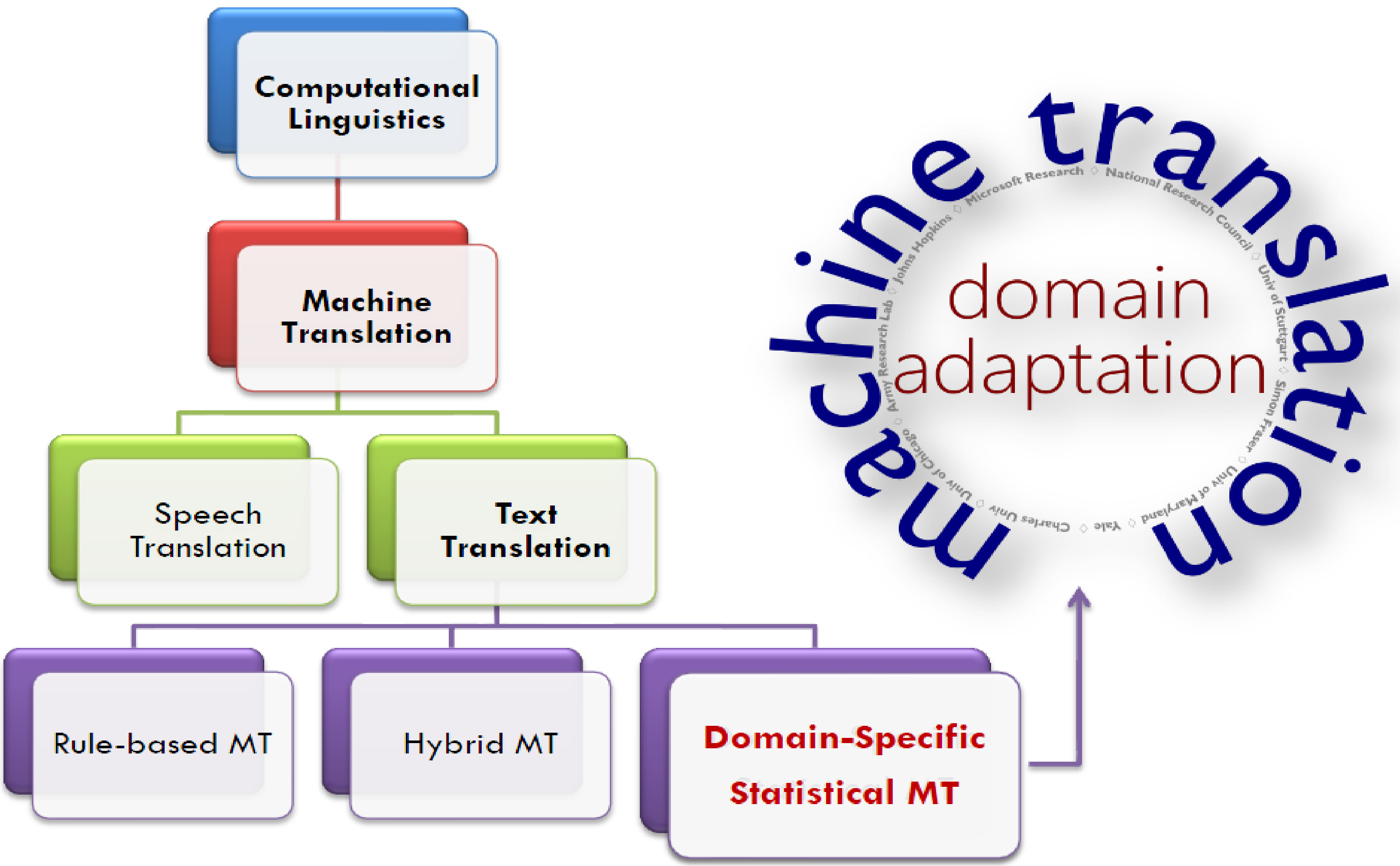


NMT (2014~)

Outline

- Introduction
  - Domain adaptation
  - Machine translation
- Domain Adaptation for SMT
  - Self-training
  - Data selection
  - Data weighting
  - Context based
  - Topic based
- Domain Adaptation for NMT
- Our work
- Conclusion && Future work

Domain adaptation for SMT



[Daniel Jurafsky et al., 2008]

Statistical Machine translation

$\mathbf{x}$ 布什 与 沙龙 举行 了 会谈

⇩

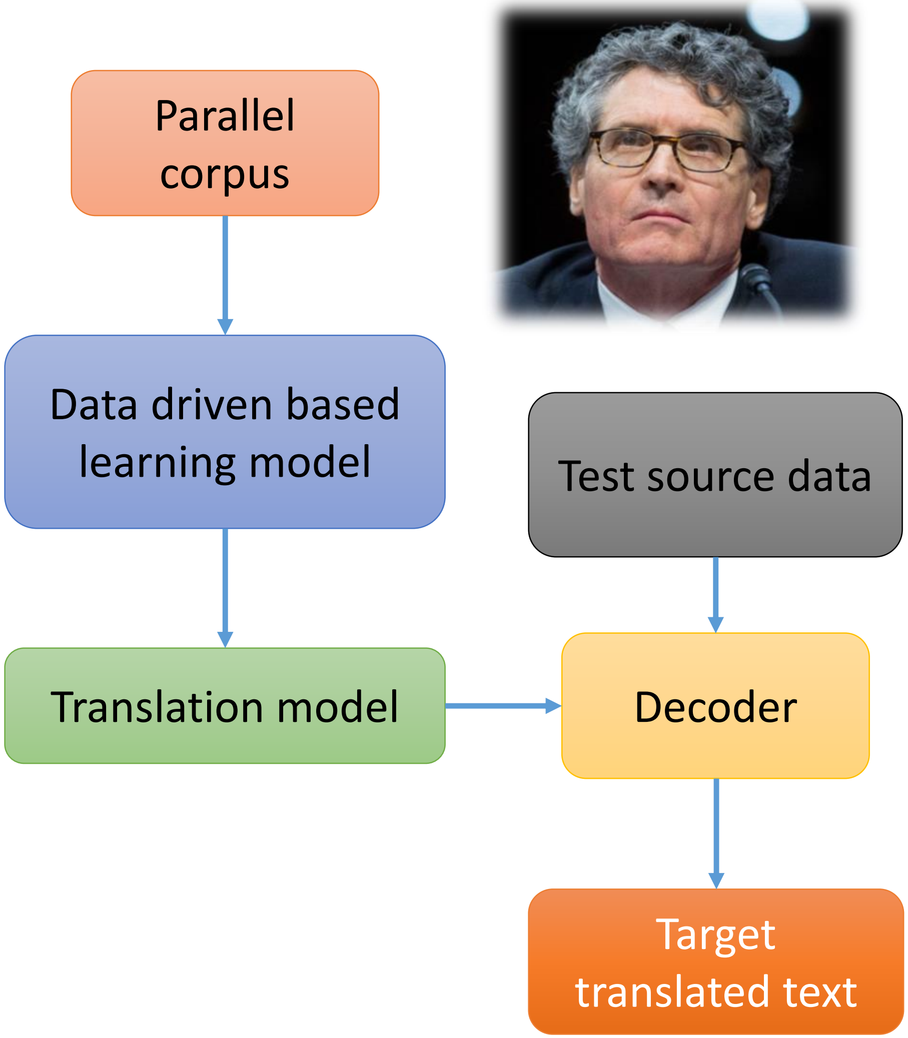
$$P(\mathbf{y}|\mathbf{x};\boldsymbol{\theta}) = \sum_{\mathbf{z}} \frac{\exp(\boldsymbol{\theta} \cdot \boldsymbol{\phi}(\mathbf{x},\mathbf{y},\mathbf{z}))}{\sum_{\mathbf{y}'}\sum_{\mathbf{z}'}\exp(\boldsymbol{\theta} \cdot \boldsymbol{\phi}(\mathbf{x},\mathbf{y}',\mathbf{z}'))}$$

⇩

$\mathbf{y}$ Bush held a talk with Sharon

[Och and Ney., 2002]

Statistical Machine translation --- Generative Model



Source sentence: $S = s_1^m = s_1 s_2 \cdots s_m$

Target sentence: $T = t_1^n = t_1 t_2 \cdots t_n$

$$P(T|S) = \frac{P(T) \times P(S|T)}{P(S)}$$

$$T' = \underset{T}{\operatorname{argmax}}\, P(T) \times P(S|T)$$

Language Model, LM

Translation Model, TM

[Brown et al., 1990, 1993]

$$T' = \underset{T}{\operatorname{argmax}} P(T|S)$$
$$= \underset{T}{\operatorname{argmax}} \frac{P(T) \times P(S|T)}{P(S)}$$
$$= \underset{T}{\operatorname{argmax}} P(T) \times P(S|T)$$

$$T' = \underset{T}{\operatorname{argmax}} P(T) \times P(T|S)$$

Translation quality $\approx$ Translation quality

$$T' = \underset{T}{\operatorname{argmax}} P(T) \times P(S|T)$$
$$T' = \underset{T}{\operatorname{argmax}} P(T) \times P(T|S)$$

$$T' = \underset{T}{\operatorname{argmax}} P(T) \times P(S|T) \times P(T|S)$$

Quality $<$ Quality

[Och and Ney., 2002]

$$T' = \underset{T}{\operatorname{argmax}} P(T|S)$$

$$= \underset{T,S_1^K}{\operatorname{argmax}} P(T, S_1^K|S)$$

$$= \underset{T,S_1^K,T_1^K,T_1^{K'}}{\operatorname{argmax}} P(S_1^K|S) \times$$ Phrase splitting model

$$P(T_1^K|S_1^K, S) \times$$ Phrase translation model

$$P\left(T_1^{K'}|T_1^K, S_1^K, S\right) \times$$ Phrase reordering model

$$P\left(T|T_1^{K'}, T_1^K, S_1^K, S\right)$$ Target language model

[Koehn, 2003]

Without Domain Adaptation

- **MT systems** make error in new domains
- OOV words are a big problem
- So are words with new senses
- Even known words with known translations can have wrong translation scores.

Word Senses vs. Domains

- **Many words have multiple senses**
- Cross-lingual mapping difficult for all contexts
- Senses are often domain – specific ?

Typical SMT vs. Domain-Specific SMT

- **Typical SMT** systems trained on a **large** and **broad** corpus (i.e., general-domain) and deal with texts with neglecting domain.
- Depends heavily upon the **quality** and **quantity** of training corpus.
- Output preserve **semantics** of the source side but lack **morphological** and **syntactic** correctness.
- Understandable translation quality.

Input:
Hollywood actor Jackie Chan has apologized over his son's arrest on drug-related charges, saying he feels "ashamed" and "sad".

Google Output:
好莱坞影星**成龙**已经**道歉**了他**儿子**的被捕与**毒品**有关的指控，说他感觉“羞耻”和“悲伤”。

Is Machine Translation good enough ?

Typical SMT vs. Domain-Specific SMT

- **Domain-Specific** SMT systems trained on a small but **relative** corpus (i.e., in-domain) and deal with texts from one specific domain.
- Consider relevance between training data and what we want to translate (test).
- Output preserve **semantics** of the source side **morphological** and **syntactic** correctness.
- Publishable quality.

Input:

本发明涉及**新的**tetramic酸型化合物，它从CCR—5活性复合物中分离出来，在控制条件下通过将生物纯的微生物培养液(球毛壳霉Kunze SCH 1705 ATCC 74489)发酵来制备复合物

ICONIC Translator Output:

Novel tetramic acid-type compounds **isolated** from a CCR-5 active complex produced by fermentation under controlled conditions of a biologically pure culture of the microorganism, Chaetomium globosum Kunze SCH 1705, ATCC 74489 ., pharmaceutical compositions containing the compounds.

- **Multi-meaning** may not coincide in bilingual environment. The English word *Mouse* refers to both animal and electronic device. While in the Chinese side, they are two words. Choosing wrong translation variants is a potential cause for **miscomprehension**.

I want to buy a mouse .

Source Side

Electronic device or animal

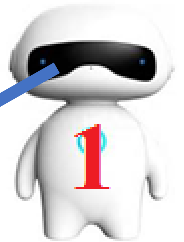
Meaning

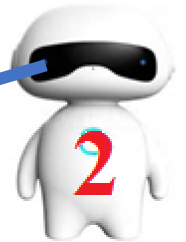
我 想 买 一张 票

我 想 买 一只 老鼠

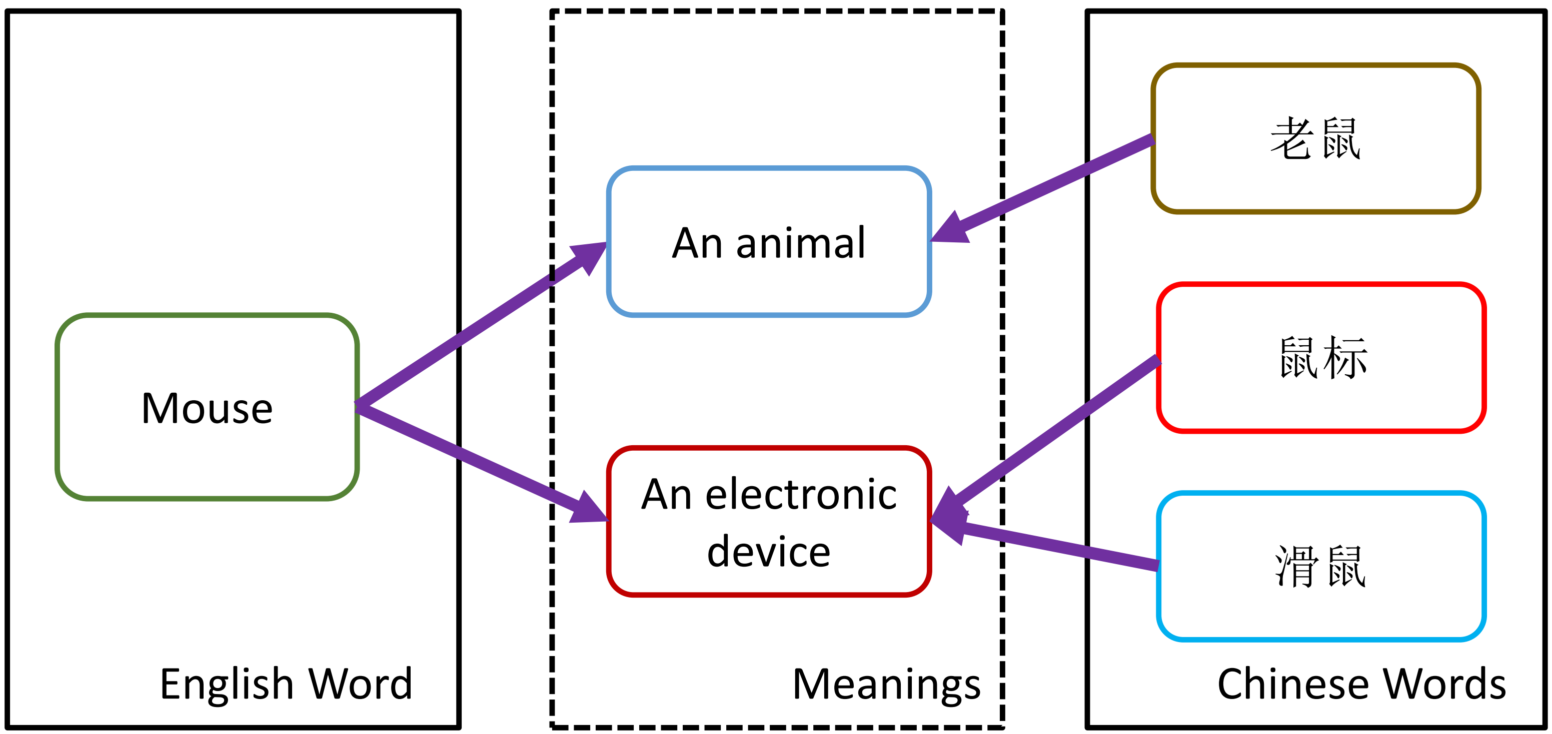
我 想 买 一个 滑鼠

Target side





- **Multi-meaning** may not coincide in bilingual environment. The English word *Mouse* refers to both animal and electronic device. While in the Chinese side, they are two words. Choosing wrong translation variants is a potential cause for **miscomprehension**.



News Domain

- Try to deliver rich information with very economical language.
- Short and simple-structure sentence make it easy to understand
- A lot of abbreviation, date, named entities.

China's Li Duihong won the women's 25-meter sport pistol Olympic gold with a total of 687.9 points early this morning Beijing time. (Guangming Daily, 1996/07/02)
我国女子运动员李对红今天在女子运动手枪决赛中，以687.9环战胜所有对手，并创造新的奥运记录。（《光明日报》1996年7月2日）

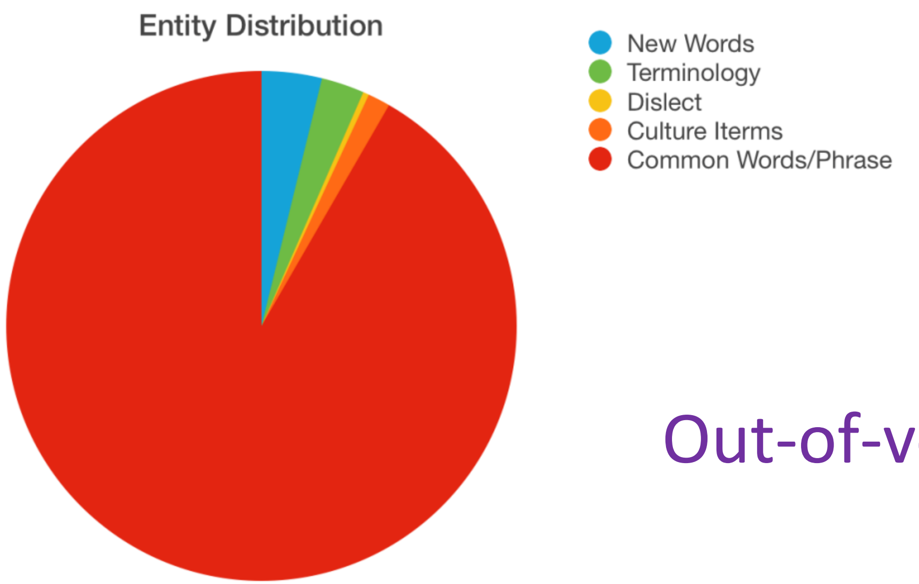
Law Domain

- Very rigorous even with duplicated terms.
- Use fewer pronouns, abbreviations etc. to avoid any ambiguity.
- High frequency words of shall, may, must, be to.
- Long sentence with long subordinate clauses.

When an international treaty that relates to a contract and **which** the People's Republic of China has concluded on participated into has provisions of the said treaty shall be applied, but with the exception of clauses to which the People's Republic of China has declared reservation.
中华人民共和国缔结或者参加的与合同有关的国际条约**同中华人民共和国**法律有不同规定的,适用该国际条约的规定。但是,**中华人民共和国声**明保留的条款除外。

Domain Specific Translation challenge 3- OOV

- Terminology: words or phrases that mainly occur in specific contexts with specific meanings.
- Variants, increasing, combination etc.



Out-of-vocabulary example

Domain adaptation for SMT

- DA can be done by model level
  - Alignment model
  - Language model
  - Translation model
  - Reordering model
- DA can also be achieved corpus level
  - Dictionary
  - Comparable corpora
  - Parallel corpora
  - Monolingual corpora
- DA approaches can be decided into:
  - Unsupervised
  - Semi-Supervised
  - Supervised

Domain adaptation for SMT

- Self-training
- Data selection
- Data weighting
- Context based
- Topic based

Domain Adaptation for Statistical Machine Translation with Monolingual Resources

Nicola Bertoldi Marcello Federico

FBK-irst - Ricerca Scientifica e Tecnologica, Italy

EACL2009, Workshop on SMT

The basic idea is that **in-domain** training data can **be exploited** to adapt all components of an already developed system. Previous work showed small performance gains by adapting from limited in-domain bilingual data.

We propose to synthesize a bilingual corpus by **translating**(with a background system) the monolingual adaptation data into the counterpart language and **train** statistical models form the synesthetic corpus.

$$S = \{(\tilde{f}, \tilde{e})\} \quad h(\tilde{f}, \tilde{e}; S)$$

$$S_I = \{(\tilde{f}, \tilde{e}) | \forall j (\tilde{f}, \tilde{e}) \in S_j\}$$

$$S_U = \{(\tilde{f}, \tilde{e}) | \exists j (\tilde{f}, \tilde{e}) \in S_j\}$$

Domain Adaptation for Statistical Machine Translation with Monolingual Resources

$$h(\tilde{f},\tilde{e};S_j)=\frac{\epsilon}{(l+1)^m}\prod_{k=1}^{m}\sum_{h=0}^{l}\emptyset(e_k|f_h)$$

Language pair	Training data		PP	OOV	BLEU	NIST	WER	PER
	TM/RM	LM						
Spanish-English	UN	UN	286	1.12	22.60	6.51	64.60	45.52
Spanish-English	UN	EP	74	0.15	27.83	7.12	60.93	45.19
Spanish-English	EP	EP	74	0.15	32.80	7.84	56.47	41.15
Spanish-English	UN	S$\bar{E}$-EP	89	0.21	23.52	6.64	63.86	47.68
Spanish-English	S$\bar{E}$-EP	S$\bar{E}$-EP	89	0.21	23.68	6.65	63.64	47.56
Spanish-English	$\bar{S}$E-EP	$\bar{S}$E-EP	74	0.15	28.10	7.18	60.86	44.85
Spanish-English	Google		Null	Null	28.60	7.55	57.38	57.38
Spanish-English	Euromatrix		Null	Null	32.99	7.86	56.36	41.12
Spanish-English	UN	UN	281	1.39	23.24	6.44	65.81	49.61

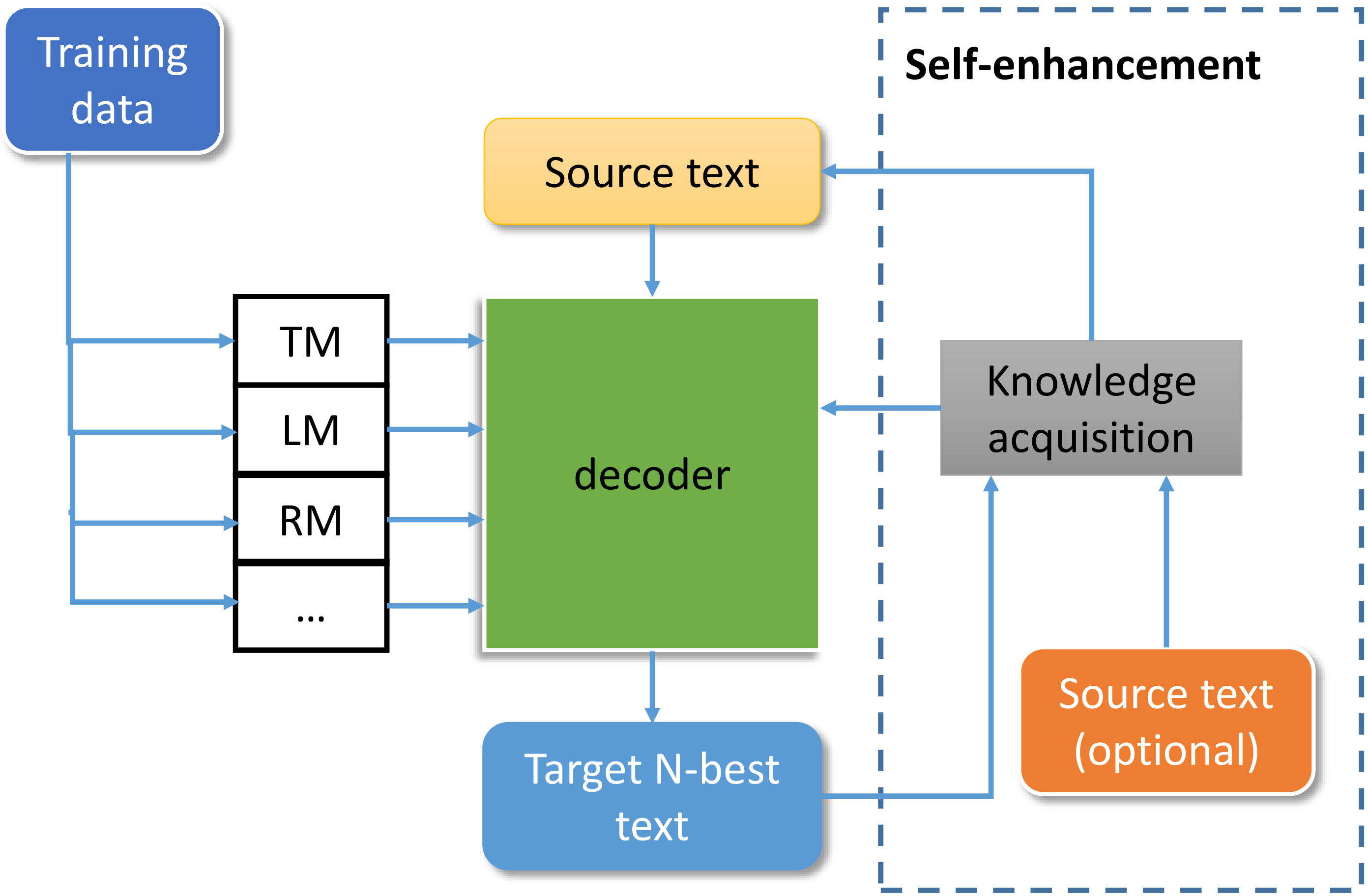
Exploiting N-best Hypotheses for SMT Self-Enhancement

Boxing Chen **Min Zhang** **Aiti Aw** **Haizhou Li**

Department of Human Language Technology,
Institute for information Research, Singapore

ACL2008

Exploiting N-best Hypotheses for SMT Self-Enhancement



$$h_{LM}(f_1^J, e_1^I) = \lambda_1 h_{TLM}(e_1^I) + \lambda_2 h_{QLM}(e_1^I)$$

$$p(\tilde{e}|\tilde{f}) = \frac{N_{train}(\tilde{f},\tilde{e}) + N_{nbest}(\tilde{f},\tilde{e})}{N_{train}(\tilde{f}) + N_{nbest}(\tilde{f})}$$

System	iteration	NIST02	NIST03	NIST05
Base	-	27.67	26.68	24.82
TM	4	27.87	26.95	25.05
LM	6	27.96	27.06	25.07
WR	6	27.99	27.04	25.11
Comb	7	28.45	27.35	25.46

Self enhancement on TM,LM,WR(word reordering model),combination

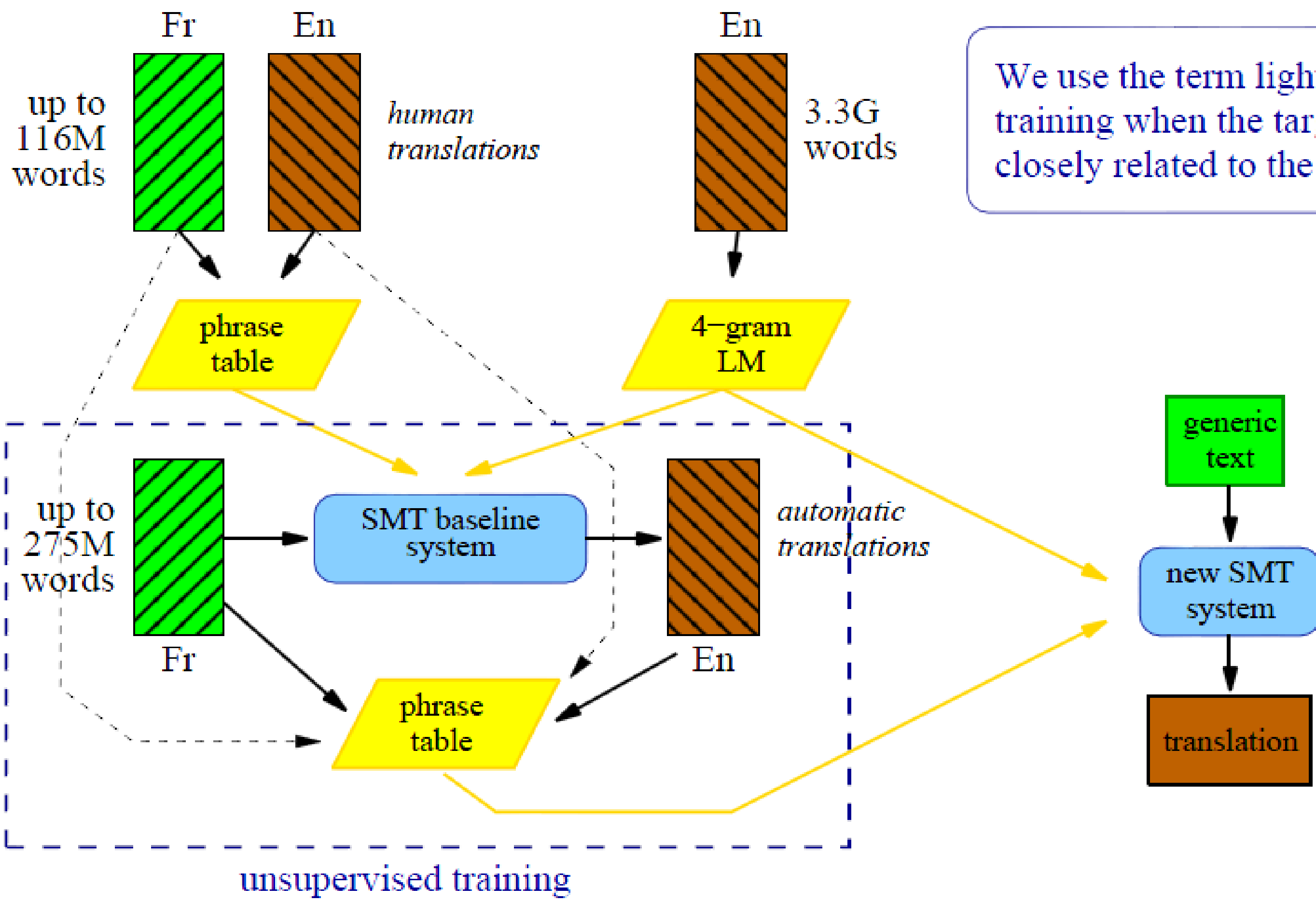
Investigations on Large-Scale Lightly-Supervised Training for Statistical Machine Translation

Holger Schwenk

LIUM, University of Le Mans, FRANCE

IWSLT2008

$$e^* = \operatorname*{argmax}_e \Pr(e|f) = \operatorname*{argmax}_e \Pr(f|e)\Pr(e) \quad e^* = \operatorname*{argmax}_e \Pr(e|f) = \operatorname*{argmax}_e \{\exp(\sum_i \lambda_i h_i(e,f))\}$$



We use the term lightly−supervised training when the target LM data is closely related to the text to be translated

Bitexts: Human-provided		Bitexts: Lightly-supervised		Total Words	BLEU score: Dev	BLEU score: Test	Phrase table Size [#entries]
News+dict	2.4M	-		2.4M	**20.44**	**20.18**	5M
News+Eparl+dict	43M			43.3M	22.17	22.35	83M
News+Eparl+Hans+dict	116M			116M	**22.69**	**22.17**	213M
Translated with the small SMT system:							
News	2.4M	afp9x	28M	2.4M	21.21	21.02	58M
			101M	2.4M	**21.23**	**21.18**	189M
		afp2x	43M	2.4M	20.98	21.01	77M
			102M	2.4M	**21.23**	**21.17**	170M
		Eparl	7M	2.4M	20.78	20.65	17M
			31M	2.4M	21.14	20.86	67M
Translated with the big SMT system:							
-		afp2x	31M	31M	22.23	22.33	55M
			112M	112M	**22.56**	**22.47**	180M
News+Eparl	42M	afp2x	77M	129M	22.65	22.44	203M
	42M		155M	197M	22.53	22.73	320M
News+Eparl+Hans	114M	afp2x	167M	281M	**22.86**	**22.80**	464M

Data selection

Selecting data suitable for the domain at hand from large **general-domain** corpora, under the **assumption** that a **general corpus** is broad enough to contain sentences that are similar to those that occur in the domain.

- Do not change the pipeline, improve the input.
- Not all sentence are equally valuable…
- For particular translation task:
  - Identify the most relevant training data
  - Build a model on only this subset
- Goal:
  - Better task-specific performance
  - Cheaper (computation, size, time)

Intelligent Selection of Language Model Training Data

Robert C. Moore William Lewis

Microsoft Research, USA

ACL2011

Intelligent Selection of Language Model Training Data

$$P(N_I|s,N) = \frac{P(s|N_I,N)P(N_I|N)}{P(s|N)}$$

$N_I \xrightarrow{\text{Subset of}} N$

$P(s|N_I,N)=P(s|N_I)$

Relationship I and N_I is $P(s|N_I)=P(s|I)$

Estimate it by training LM on I

$$P(N_I|s,N) = \frac{P(s|I)P(N_I|N)}{P(s|N)}$$

Estimate it by training LM on N

$H_I(s)$ Per word corss-entropy according to LM on I, text segment s drown from N

$H_N(s)$ Per word corss-entropy according to LM on N

Partition N into segments (sentences), according to $H_I(s)$-$H_N(s)$ score segments.

$$\log(P(s|I)) - \log(P(s|N)) \approx H_I(s)\text{-}H_N(s)$$

Intelligent Selection of Language Model Training Data

Corpus	Sentence country	Token count
Gigaword	133,310,562	3,445,946,266
Europarl train	1,651,392	48,230,859
Europarl test	2,000	55,566

Selection Method	Original LM PPL	Modified LM PPL
In-domain cross-entropy scoring	124.4	124.8
Klakow's method	110.5	110.8
Cross-entropy difference scoring	100.7	101.9

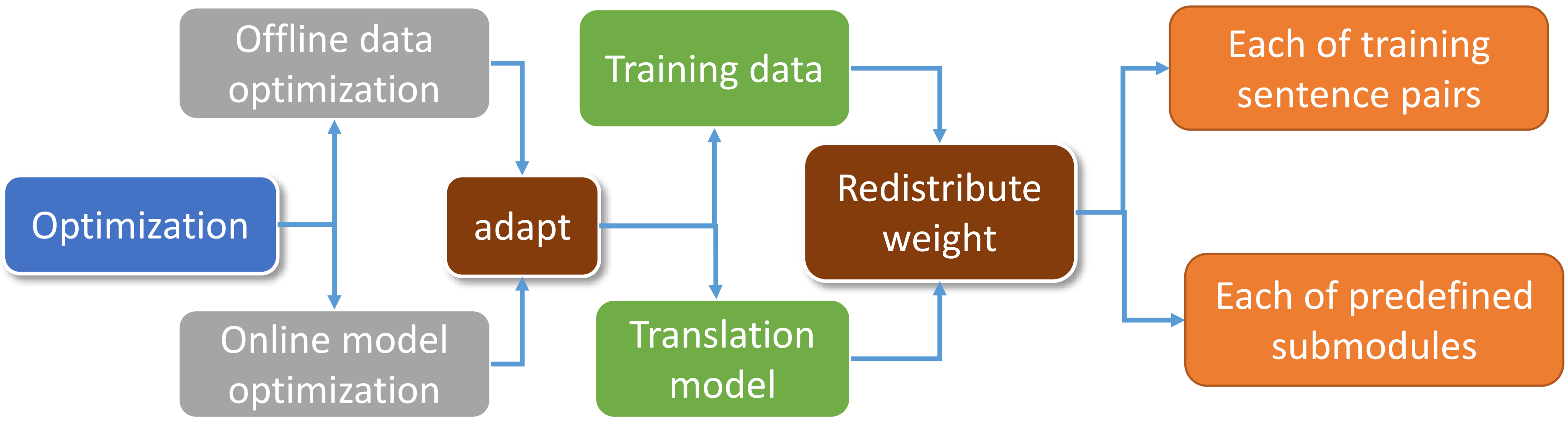
Improving Statistical Machine Translation Performance by Training Data Selection and Optimization

Yajuan Lü, Jin Huang and Qun Liu

Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences

EMNLP2007

Improving SMT Performance by Training Data Selection and Optimization



Similar data selection by TF-IDF

$$D_i = (W_{i1}, W_{i2}, \cdots, W_{in})$$

Vocabulary size = n

$$W_{ij} = tf_{ij} \times \log(idf_j)$$

Online model weighting

$$\hat{p}(e|c) = p_0(e|c)^{\delta_0} \times \prod_{i=1}^{M} p_i(e|c)^{\delta_i}$$

$$\hat{e} = \operatorname*{argmax}_{e}(\delta_0 \log(p_0(e|c)) + \sum_{i=1}^{M} \delta_i \log(p_i(e|c)))$$

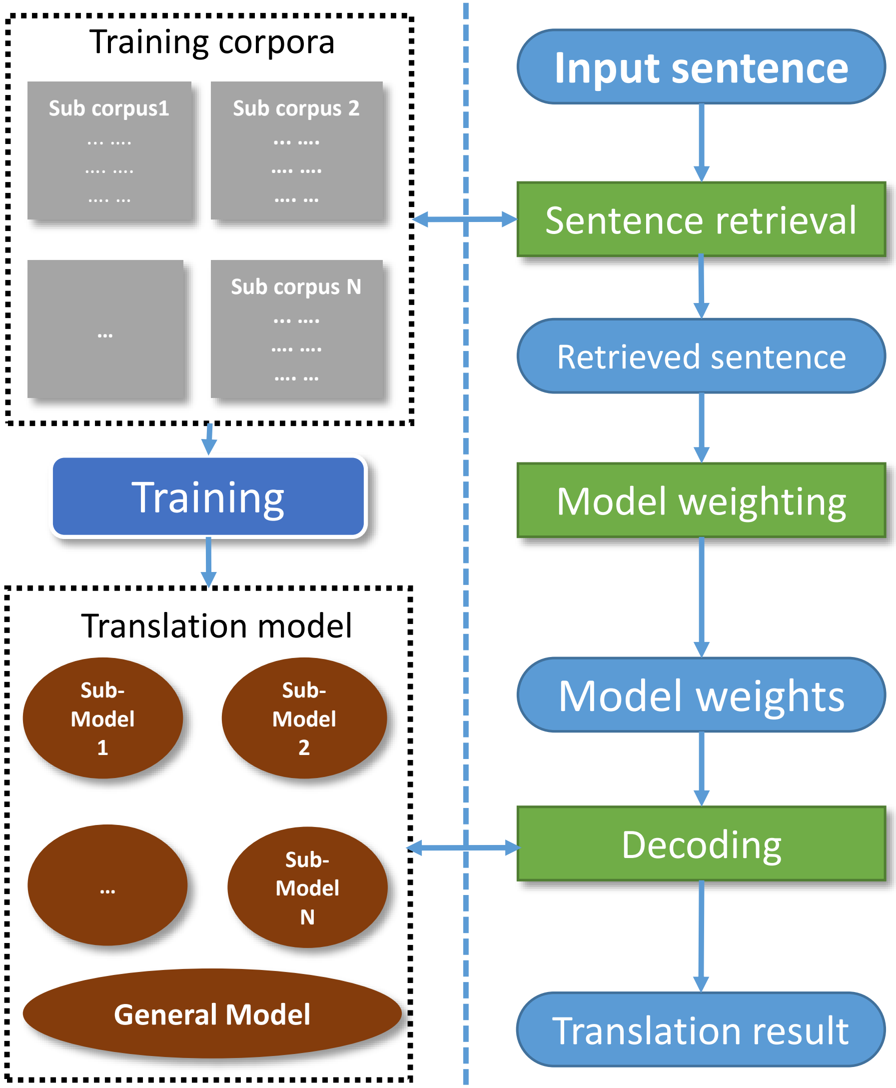
p_0 and p_i are general model and submodule

δ_0 and δ_i *are weights*

Improving Statistical Machine Translation Performance by Training Data Selection and Optimization

Training procedure

Translation procedure



Offline data optimization result

System	Distinct pairs	Blue on TopN	Blue on TopN+
Baseline	600000	0.2363	0.2363
Top 100+	600000	0.2306	0.2387
Top 200+	600000	0.2360	0.2443
Top 500+	600000	0.2415	0.2461
Top 1000+	600000	0.2463	0.2431
Top 2000+	600000	0.2351	0.2355

Online model optimization result

System / Test data	S_1	S_2	S_3	S_4
FBIS-part	0.1090	0.1090	0.1089	0.1089
HK_Hans_part	0.0906	0.0903	0.0902	0.0902
HK_News_part	0.0952	0.0950	0.0933	0.0934
MT05_part	0.1119	0.1123	0.1149	0.1151
Whole_test_set	0.1034	0.1034	0.1038	0.1038

Domain Adaptation via Pseudo In-Domain Data Selection

Amittai Axelrod, Xiaodong He, Jianfeng Gao

University of Washington && Microsoft Research

EMNLP2011

Perplexity-based model, which employs *n*-gram in-domain language models to score the perplexity of each sentence in general-domain corpus.

Cross-entropy is the average of the negative logarithm of the word probabilities.

$$\mathrm{H}(\mathrm{p},\mathrm{q}) = -\sum_{i=1}^{n} p(w_i)\log q(w_i) = -\frac{1}{N}\sum_{i=1}^{n} log q(w_i)$$

Perplexity *pp* can be simply transformed with a base *b* with respect to which the cross-entropy is measured.

$$\mathrm{pp} = b^{\mathrm{H}(\mathrm{p},\mathrm{q})}$$

Perplexity and cross-entropy are **monotonically related**

The first **basic** one $H_{I-\mathrm{src}}(\mathrm{x})$

The second is called **Moore-Lewis** $H_{I-\mathrm{src}}(\mathrm{x}) - H_{O-\mathrm{src}}(\mathrm{x})$

which tries to select the sentences that are more similar to in-domain but different to out-of-domain.

The third is **modified Moore-Lewis**

$$[H_{I-\mathrm{src}}(\mathrm{x}) - H_{O-\mathrm{src}}(\mathrm{x})] + [H_{I-tgt}(\mathrm{x}) - H_{O-tgt}(\mathrm{x})]$$

which considers both source and target language

Domain Adaptation via Pseudo In-Domain Data Selection

Concatenating in-domain and pseudo [single Model]

Method	sentences	Dev	Test
IWSLT	30K	45.43	37.17
Bilingual M-L	35k	39.59	42.31
Bilingual M-L	70k	40.84	42.29
Bilingual M-L	150k	42.64	42.22
IWSLT+Bilingual M-L	35k	47.71	41.78
IWSLT+Bilingual M-L	70k	47.80	42.30
IWSLT+Bilingual M-L	150k	48.44	42.01

Concatenating in-domain and pseudo [together]

Method	Dev	Test
IWSLT	45.43	37.17
General	42.62	40.51
Both IWSLT, General	49.13	42.50
IWSLT,Bilingual M-L 35k	48.51	40.38
IWSLT,Bilingual M-L 70k	49.65	40.45
IWSLT,Bilingual M-L 150k	49.50	41.40
IWSLT,IWSLT+Bilingual M-L 35k	48.85	39.82
IWSLT,IWSLT+Bilingual M-L 70k	49.10	43.00
IWSLT,IWSLT+Bilingual M-L 150k	49.80	43.23

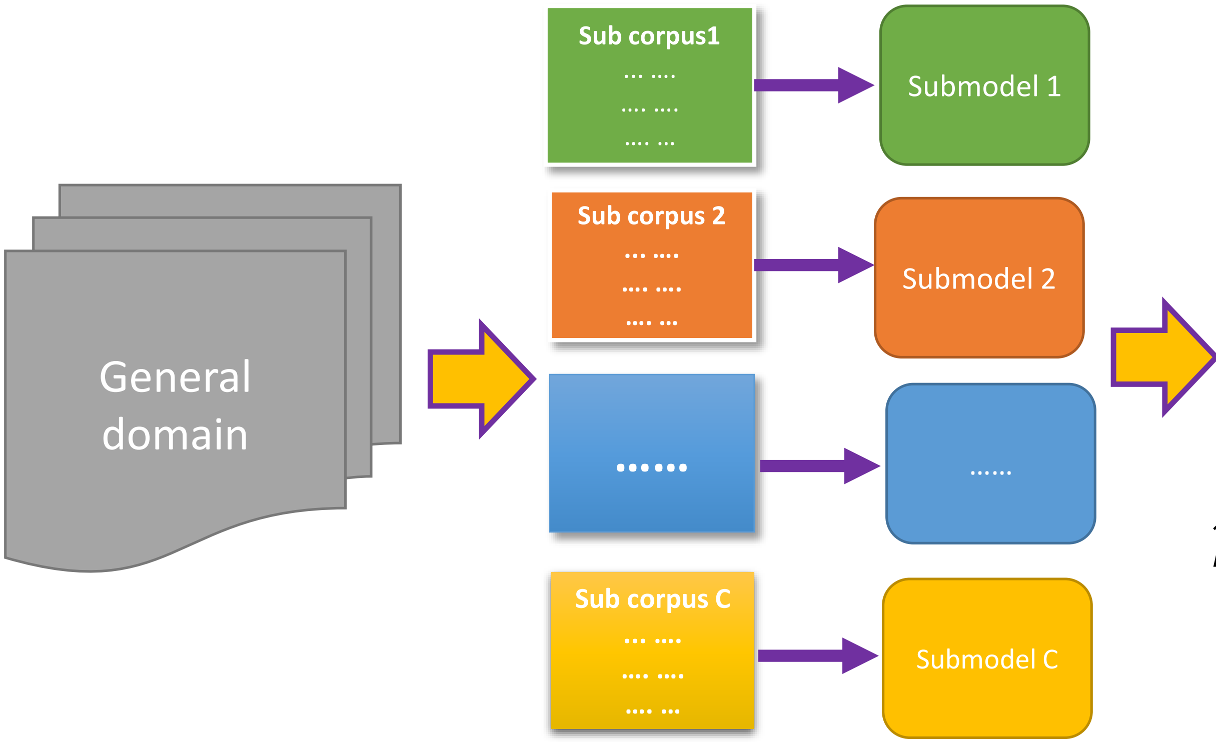
Mixture-Model Adaptation for SMT

George Foster and **Roland Kuhn**

National Research Council Canada

ACL2007

Mixture-Model Adaptation for SMT



$$\lambda_{\mathrm{c}} = \frac{\mathrm{d}_{i,\mathrm{c}}}{\sum_{c'} \mathrm{d}_{i,c'}}$$

$$p(x|h) = \sum_{c} \lambda_{\mathrm{c}} \, p_{\mathrm{c}}(x|h)$$

Distance Metrics for Weighting ： tf/idf , LSA, perplexity, EM

Mixture-Model Adaptation for SMT

Corpora

Role	Corpus	Genres	Sent
train	FBIS04	nw	182k
	HK Hans	proceedings	1,375k
	HK Laws	legal	475k
	HK News	Press release	740k
	Newswire	nw	26k
	Sinorama	news mag	366k
	UN	Proceedings	4,979k
dev	NIST04-nw	nw	901
	NIST04-mix	nw,sp,ed	889
test	NIST05	nw	1,082
	NIST06-Gale	nw,ng,bn,bc	2,276
	NIST06-NIST	nw,ng,bn	1,664

Distance matrices for linear combination on dev

Metric	Src LM	Text LM	Trg LM	Text LM
tf/idf	31.3	31.3	31.1	31.1
LSA	31.5	31.6		
Perplexity	31.6	31.3	31.7	31.5
EM	31.7	31.6	32.1	31.3

Source granularity on dynamic adaptation

Granularity	dev	test		
	Nist04-mix	nist05	Nist06-nist	Nist06-gale
Baseline	31.9	30.4	27.6	12.9
File	32.4	30.8	28.6	13.4
Genre	32.5	31.1	28.9	12.2
Document	32.9	30.9	28.6	12.4

Perplexity Minimization for Translation Model Domain Adaptation in Statistical Machine Translation

Rico Sennrich

Institute of Computational Linguistics, University of Zurich

EMNLP2012

Perplexity Minimization for Translation Model DA in SMT

A **weighted combination** can control the contribution of the **out-of-domain** corpus on the probability distribution, and thus limit the **ambiguity** problem.

A **weighted combination** eliminates the need for **data selection**, offering a robust baseline for domain-specific machine translation.

Aim to adapt all features: $p(\bar{t}|\bar{s})$ $p(\bar{s}|\bar{t})$ $lex(\bar{t}|\bar{s})$ $lex(\bar{s}|\bar{t})$

Linear interpolation model: $p(x|y;\lambda) = \sum_{i=1}^{n} \lambda_i p_i(x|y)$ $\sum_{i=1}^{n} \lambda_i = 1$

Weighted counts: $p(x|y) = \frac{c(x,y)}{c(y)} = \frac{c(x,y)}{\sum_{x'} c(x',y)}$ $p(x|y;\lambda) = \frac{\sum_{i=1}^{n} \lambda_i c_i(x,y)}{\sum_{i=1}^{n} \sum_{x'} \lambda_i c_i(x,y)}$

Perplexity minimization:

$$H(p) = -\sum_{x,y} \tilde{p}(x,y) log_2 p(x|y)$$

$$\hat{\lambda} = \underset{\lambda}{\operatorname{argmin}} - \sum_{x,y} \tilde{p}(x,y) log_2 p(x|y;\lambda)$$

Perplexity Minimization for Translation Model DA in SMT

System	out-of-domain LM		adapted LM			
	full IN TM		full IN TM		small IN TM	
	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR
in-domain	30.4	30.7	33.4	31.7	29.7	28.6
out-of-domain	24.3	28.0	28.9	30.2	28.9	30.2
counts (concatenation)	30.3	31.2	33.6	32.4	31.3	31.3
binary in/out						
weighted counts	31.0	31.6	33.8	32.4	31.5	31.3
linear interpolation (naive)	30.8	31.4	33.7	32.4	31.9	31.3
linear interpolation (modified)	30.8	31.5	33.7	32.4	31.7	31.2
alternative paths	30.8	31.3	33.2	32.4	29.8	30.7
10 models						
weighted counts	31.0	31.5	33.5	32.3	31.8	31.5
linear interpolation (naive)	30.9	31.4	33.8	32.4	31.9	31.3
linear interpolation (modified)	31.0	31.6	33.8	32.5	32.1	31.5
alternative paths	25.9	29.2	24.3	29.1	29.8	30.9

Context Adaptation in Statistical Machine Translation Using Models with Exponentially Decaying Cache

JÖrg Tiedemann

Department of Linguistics and Philology, Uppsala University, Uppsala/Sweden

ACL2010

Context Adaptation in SMT Using Models with Exponentially Decaying Cache

Mix a **large global** (static) LM **with** a small local(Dynamic model) estimated from **recent items in the history of the input stream.**

"They may also have **episodes** of depression . Abilify is used to treat moderate to severe manic **episodes** and to prevent manic **episodes** in patients who have responded to the **medicine** in the past . The solution for injection is used
for the rapid control of agitation or disturbed behavior when taking the **medicine** by mouth is not appropriate .The **medicine** can only be obtained with a prescription ."

The 10 commandments

To some land flowing with milk and honey!
Till ett land fullt av mjölk och **honung**.

I've never tasted honey.
Jag har aldrig smakat ho-**nung**.
...

Kerd ma lui

Mari honey ...
Mari, **gumman** ...

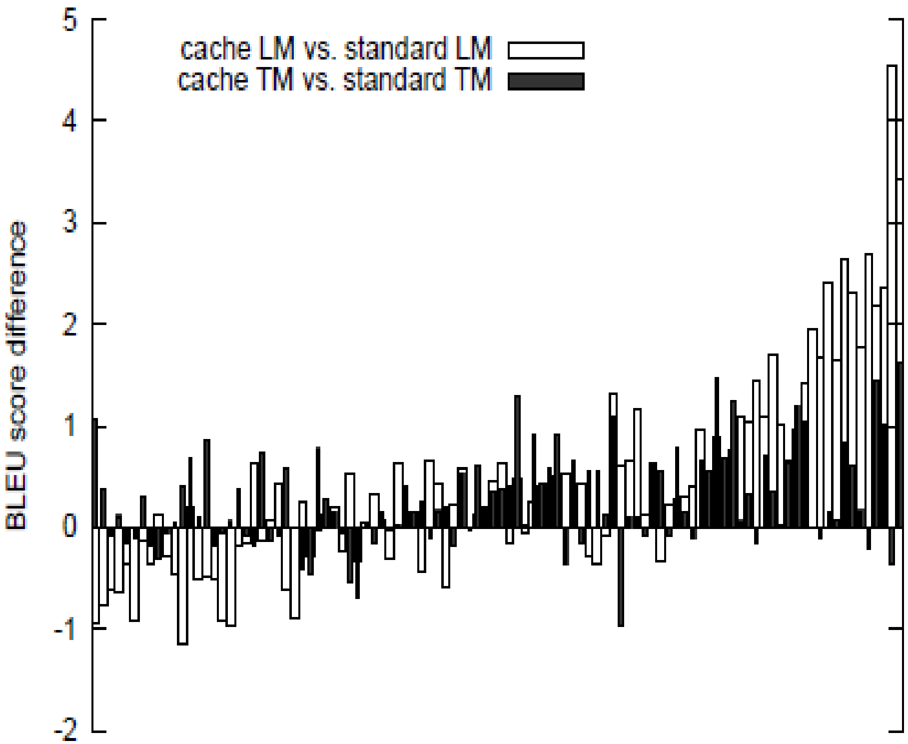
Sweetheart, where are you going?
Älskling, var ska du?
...
Who was that, honey?
Vem var det, **gumman**?

$$P(w_n|history) = (1-\lambda)P_{n-gram}(w_n|history) + \lambda P_{cache}(w_n|history)$$

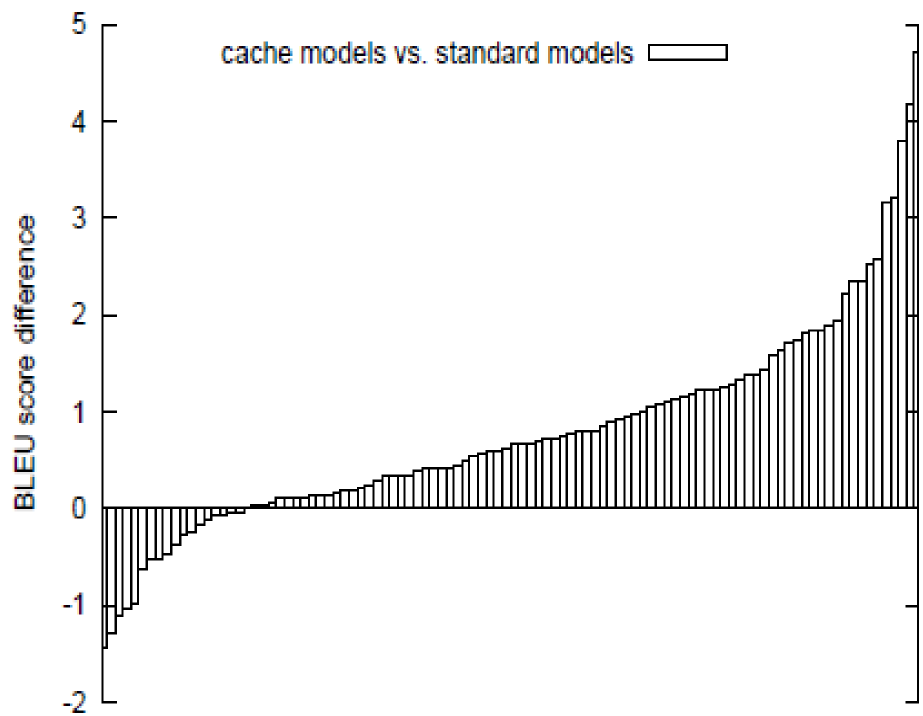
$$P_{cache}(w_n|w_{n-k} \dots w_{n-1}) \approx \frac{1}{Z}\sum_{i=n-k}^{n-1} I(w_n = w_n)e^{-\alpha(n-i)}$$

Context Adaptation in SMT Using Models with Exponentially Decaying Cache

$$\emptyset_{cache}(e_n|f_n) = \frac{\sum_{i=1}^{K} I(<e_n, f_n> = <e_i, f_i>) * e^{-\alpha i}}{\sum_{i=1}^{K} I(f_n = f_i)}$$



standard model and models with cache(102)



standard model and a model with cache (LM,TM)

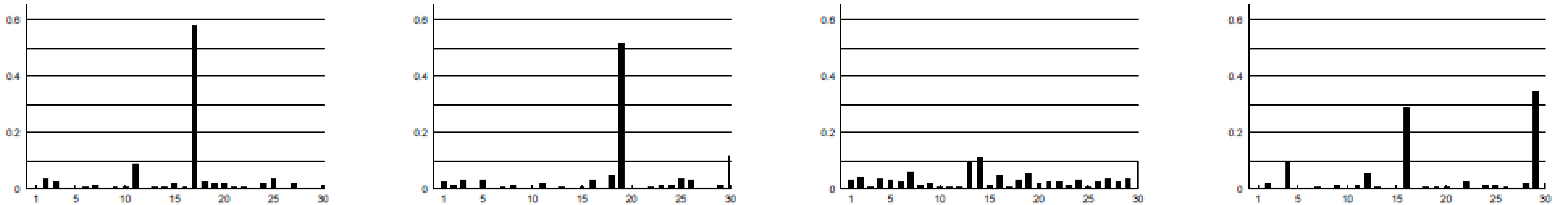
A Topic Similarity Model for Hierarchical Phrase-based Translation

Xinyan Xiao Deyi Xiong Min Zhang Qun Liu Shouxun Lin

Institute of Computing Technology, Chinese Academy of Sciences

ACL2012

A Topic Similarity Model for Hierarchical Phrase-based Translation



(a) 作战 能力 ⇒ operational capability (b) 给予 X_1 ⇒ grands X_1 (c) 给予 X_1 ⇒ give X_1 (d) X_1 举行 会谈 X_2 ⇒ held talks X_1 X_2

$$Similarity\ (P(z|d), P(z|r)) = \sum_{k=1}^{K} (\sqrt{P(z=k|d)} - \sqrt{P(z=k|r)})^2$$

$$Similarity\ P(z|r) = -\sum_{k=1}^{K} P(z=k|r) \times \log(P(z=k|r))$$

$$P(z=k|r) = \frac{\sum_{I \in \mathrm{I}} c \times P(z=k|d)}{\sum_{k'}^{K} \sum_{I \in \mathrm{I}} c \times P(z=k'|d)}$$

$$\sum_{(z_f, z_e, a)} \sum_{(i,j) \in a} \delta(z_{f_i}, k_f) * \delta(z_{e_i}, k_e)$$

$$T(P(z_e|r)) = P(z_e|r) \otimes M_{K_e \times K_f}$$

Decoding

$$Similarity\ (P(z_f|d), P(z_f|r))$$

$$Similarity\ (P(z_f|d), TP(z_e|r))$$

$$Sensitivity\ (P(z_f|r)$$

$$Sensitivity(TP(z_e|r))$$

A Topic Similarity Model for Hierarchical Phrase-based Translation

BLEU and speed	System	MT06	MT08	Avgerage	Speed
hierarchical system	Baseline	30.20	21.93	26.07	12.6
topic-specific lexicon	TopicLex	30.65	22.29	26.47	3.3
similarity by source	SimSrc	30.41	22.69	26.55	11.5
similarity by target	SimTgt	30.51	22.39	26.45	11.7
two similarity	SimSrc+SimTgt	30.73	22.69	26.71	11.2
sensitivity features	Sim+Sen	30.95	22.92	26.94	10.2

Percentage of topic-sensitive rules

Type	Count	Src%	Tgt%
Phrase-rule	3.9M	83.4	84.4
Monotone-rule	19.2M	85.3	86.1
Reordering –rule	5.7M	85.9	86.8
All-rule	28.8M	85.1	86.0

Topic model on three types of rules

Type	MT06	MT08	Avg
Baseline	30.20	21.93	26.07
Phrase-rule	30.53	22.29	26.41
Monotone-rule	30.72	22.62	26.67
Reordering –rule	30.31	22.40	26.36
All-rule	30.95	22.92	26.94

Outline

- Introduction
  - Domain adaptation
  - Machine translation
- Domain Adaptation for SMT
  - Self-training
  - Data selection
  - Data weighting
  - Context based
  - Topic based
- Domain Adaptation for NMT
- Our work
- Conclusion && Future work

Neural Machine translation

$\mathbf{x}$ 布什 与 沙龙 举行 了 会谈

⇩

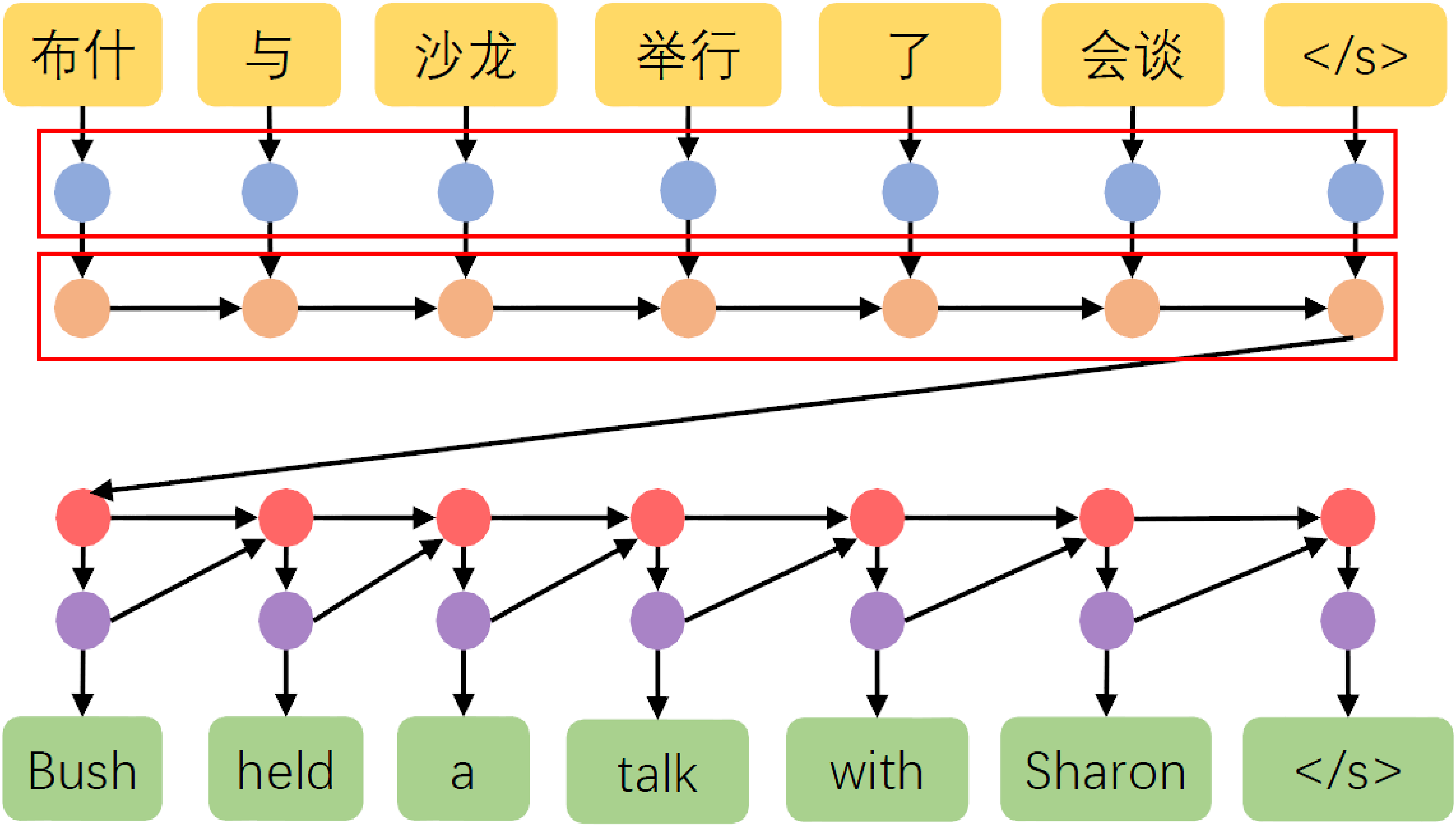
$$P(\mathbf{y}|\mathbf{x};\boldsymbol{\theta}) = \prod_{n=1}^{N} P(\mathbf{y}_n|\mathbf{x},\mathbf{y}_{<n};\boldsymbol{\theta})$$

⇩

$\mathbf{y}$ Bush held a talk with Sharon

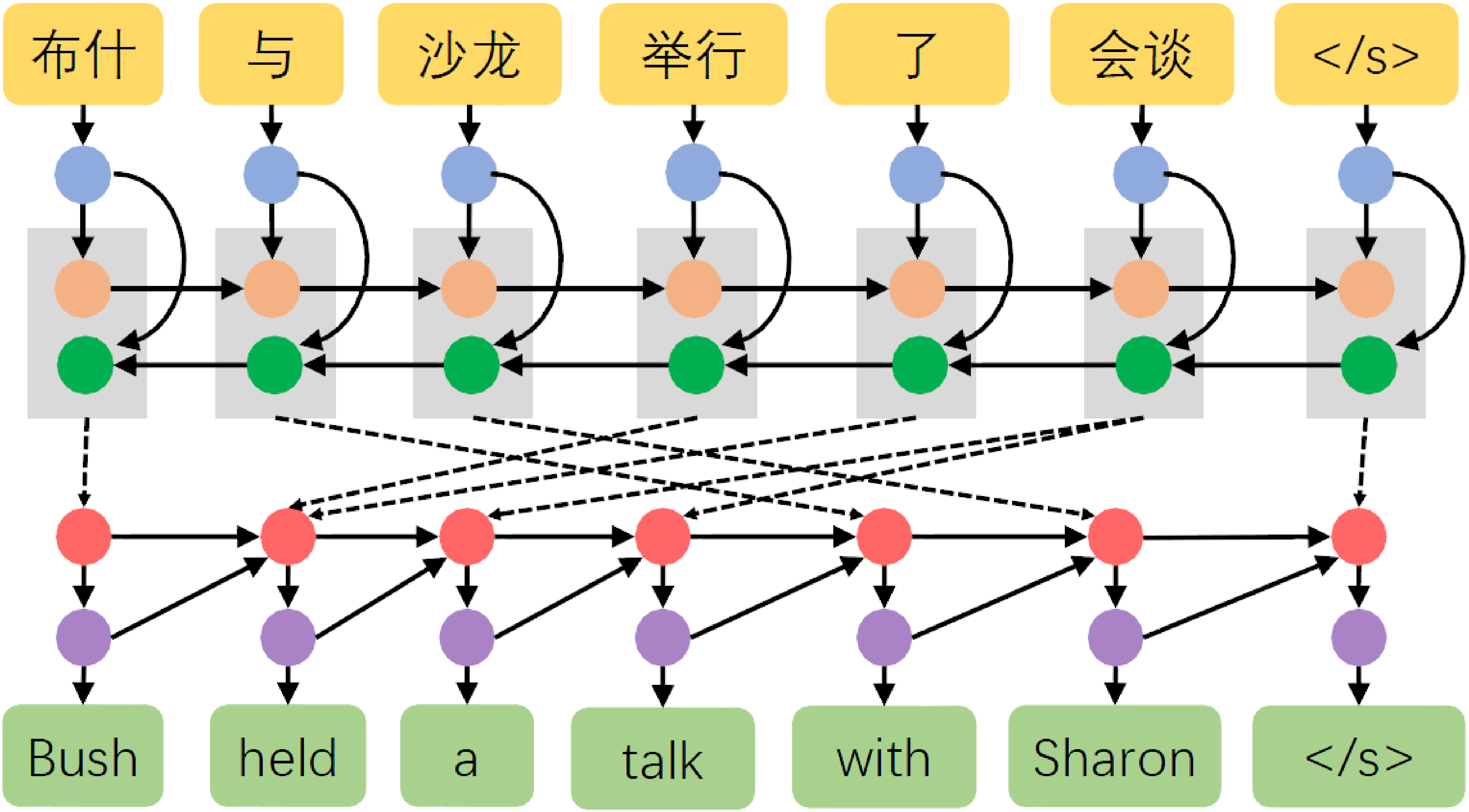
[Sutskever et al., 2014]

Neural Machine translation



[Sutskever et al., 2014]

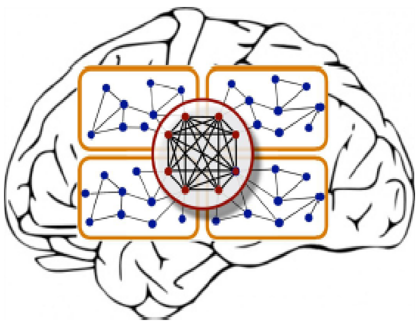
Neural Machine translation

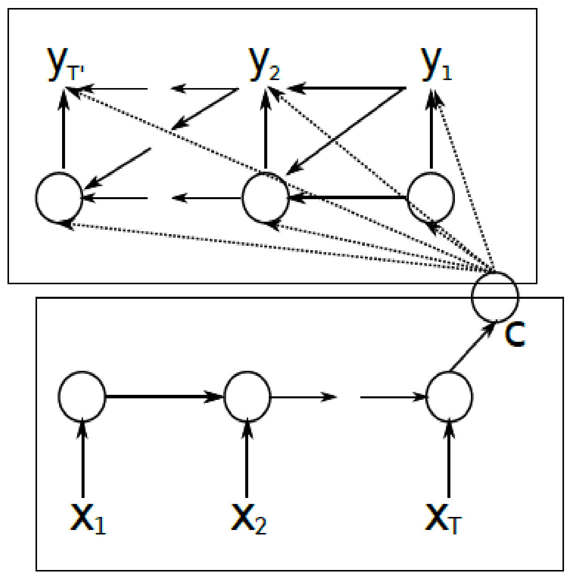


[Bahdanau et al., 2015]

Neural Machine translation

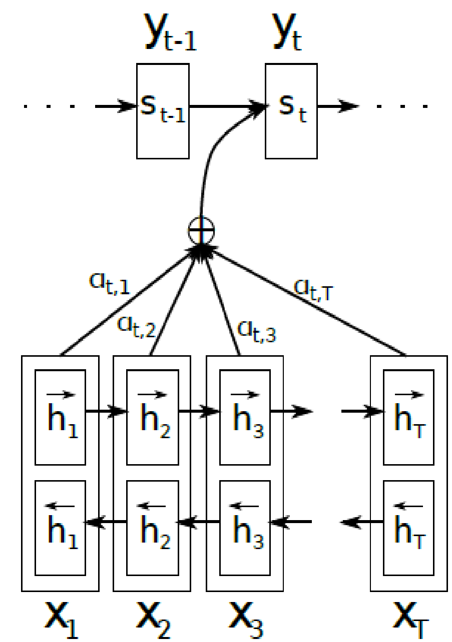
our focus





Encoder-Decoder NMT

Cho et al. (2014)



Attention-based NMT

Bahdanau et al. (2015)

Topic-Informed Neural Machine Translation

Jian Zhang, Liangyou Li, Andy Way, Qun Liu

ADAPT Centre, School of Computing, Dublin City University, Ireland

COLING2016

Topic-Informed Neural Machine Translation

Commercial analysis and market stock prices on Britain' s biggest bank .

[⋯, Financial topic, ⋯]

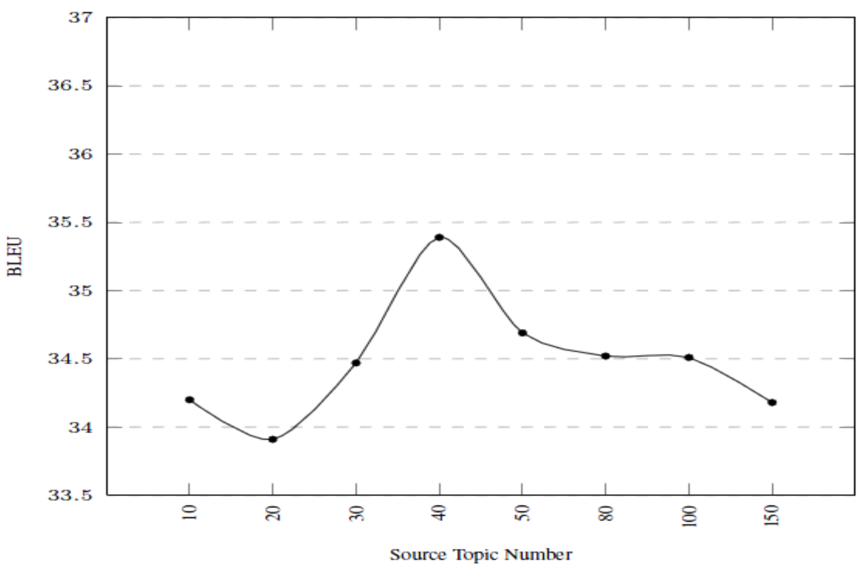
$$h_j = g(\mathrm{t}_{j-1}, h_{j-1}, \mathrm{c})$$

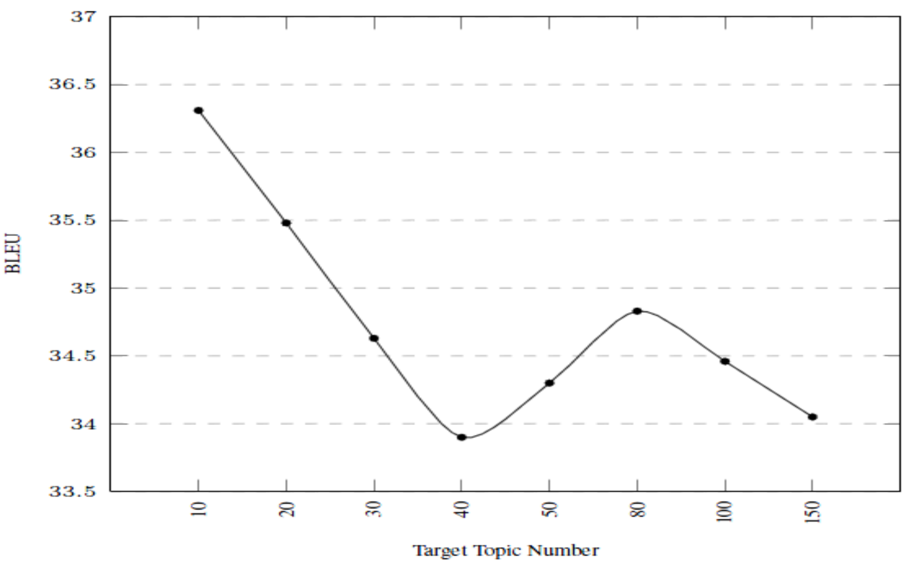
Topic-informed source context vector $topic\_c_j = \sum_{i=1}^{m} \alpha_{ij} [h_i, \beta_i^S]$

$$h_j = g(\mathrm{t}_{j-1}, h_{j-1}, topic\_c_j) \qquad h_j = g(\mathrm{t}_{j-1}, h_{j-1}, \mathrm{c}, h_{j-1}^{\beta^T})$$

$$h_j = g(\mathrm{t}_{j-1}, h_{j-1}, topic\_c_j, h_{j-1}^{\beta^T})$$

Topic-Informed Neural Machine Translation





Systems	NIST02(dev)	NIST04(test)	NIST05(test)
SMT	33.42	32.36	30.11
NMT	34.33	34.76	31.12
Source Topic-Informed NMT(40)	35.39	35.17+	31.95++
Target Topic-Informed NMT(10)	36.31	35.43++	32.50++
Topic-Informed NMT(40,10)	34.86	35.91++	32.79++

Sentence Embedding for Neural Machine Translation Domain Adaptation

Rui Wang, Andrew Finch, Masao Utiyama and Eiichiro Sumita

National Institute of Information and Communications Technology (NICT), Kyoto, Japan

ACL2017

Sentence Embedding for NMT Domain Adaptation

Source sentence as a **fixed** length vector H In-domain F_{in} out-domain F_{out}

French-to-English NMT system N_{FE} trained on F_{in} and F_{out} together.

$$s_{init}(X) = \tanh\left(W \frac{\sum_{i=1}^{T_x} h_i}{T_x} + b \right), h_i \in H$$

Vector centers $C_{F_{in}}$ $C_{F_{out}}$

Sentence embedding $v_f = s_{init}(f)$ Euclidean distance $d(v_f, C_{F_{in}}), d(v_f, C_{F_{out}})$

$$C_{F_{in}} = \frac{\sum_{f \in F_{in}} v_f}{|F_{in}|}$$

$$C_{F_{out}} = \frac{\sum_{f \in F_{out}} v_f}{|F_{out}|}$$

Classify each sentence via difference: δ

$$\delta_f = d(v_f, C_{F_{in}}) - d(v_f, C_{F_{out}})$$

$$\delta_e = d(v_e, C_{F_{in}}) - d(v_e, C_{F_{out}})$$

$$\delta_{fe} = \delta_f + \delta_e$$

Sentence Embedding for NMT Domain Adaptation

IWSLT : EN-FR

Method	Sent.	SMT tst10	SMT tst11	NMT tst10	NMT Tst11
in	178.1K	31.06	32.50	29.23	30.00
out	17.7M	30.04	29.29	27.30	28.48
Int+out	17.9M	30.00	30.26	28.89	28.55
Random	5.5M	31.22	33.85	30.53	32.37
Luong	17.9M	N/A	N/A	32.21	35.03
Axelrod	9.0M	32.06	34.81	32.26	35.54
Chen	7.3M	31.42	33.78	30.32	33.81
δ_f	7.3M	31.46	33.13	32.13	34.81
δ_e	3.7M	**32.08**	35.94	32.84	36.56
δ_{fe}	5.5M	31.79	**35.66**	32.67	36.64
δ_f+fur	7.3M	N/A	N/A	34.04	37.18
δ_e+fur	3.7M	N/A	N/A	33.88	38.04
δ_{fe}+fur	5.5M	N/A	N/A	**34.52**	**39.02**

NIST : ZH-EN

Method	Sent.	SMT MT05	SMT MT06	NMT MT05	NMT MT06
in	430.8K	29.66	30.73	27.28	26.82
out	8.8M	29.61	30.13	28.67	27.79
Int+out	9.3M	30.23	30.11	28.91	28.22
Random	5.7M	29.90	30.18	28.02	27.49
Luong	9.3M	N/A	N/A	29.91	29.61
Axelrod	2.2M	30.52	30.96	28.41	28.75
Chen	4.8M	30.64	31.05	28.39	28.06
δ_f	4.8M	30.90	**31.96**	29.21	30.14
δ_e	2.2M	**30.94**	31.33	30.00	30.63
δ_{fe}	5.7M	30.72	31.33	30.13	31.07
δ_f+fur	4.8M	N/A	N/A	30.80	31.54
δ_e+fur	2.2M	N/A	N/A	30.49	31.13
δ_{fe}+fur	5.7M	N/A	N/A	**31.35**	**31.80**

Outline

- Introduction
  - Domain adaptation
  - Machine translation
- Domain Adaptation for SMT
  - Self-training
  - Data selection
  - Data weighting
  - Context based
  - Topic based
- Domain Adaptation for NMT
- Our work
- Conclusion && Future work

Our work

[This slide intentionally left blank]

Outline

- Introduction
  - Domain adaptation
  - Machine translation
- Domain Adaptation for SMT
  - Self-training
  - Data selection
  - Data weighting
  - Context based
  - Topic based
- Domain Adaptation for NMT
- Our work
- Conclusion && Future work

Conclusion

- As SMT is **corpus-driven**, domain-specificity of training data with respect to the test data is a significant factor that we cannot ignore.
- There is a **mismatch** between the domain of available training data and the target domain.
- Unfortunately, the training resources in **specific domains** are usually relatively **scarce**.

In such scenarios, various **domain adaptation** techniques are employed to improve domain-specific translation quality by leveraging general-domain data.

Conclusion

- **VSM-based**: cosine tf-idf
- **Perplexity-based**: basic cross-entropy, Moore-Lewis and modified Moore-Lewis.
- **String-difference**: edit-distance.
- **Combination**: Corpus-level and Model-level

Above methods only consider **word itself** (surface information).

- Languages have a larger set of different words leads to **sparsity** problems.
- Weak at capturing **language style**, sentence **structure**, **sematic** information.

- **VSM-based**: cosine tf-idf
- **Perplexity-based**: basic cross-entropy, Moore-Lewis and modified Moore-Lewis.
- **String-difference**: edit-distance.
- **Combination**: Corpus-level and Model-level

Above methods only consider **word itself** (surface information).

- Languages have a larger set of different words leads to **sparsity** problems.
- Weak at capturing **language style**, sentence **structure**, **sematic** information.

Future work

- Data Selection
  - Graphical model and label propagation
  - Neural language model
- Sentence embedding
- Context based
- Topic info
- Multi – domain
- Corpus
- Model
  - LM
  - TM


THANK YOU