Semantic, Stylistic & Other Data Divergences in Neural Machine Translation

Marine Carpuat
marine@cs.umd.edu

COMPUTER SCIENCE
UNIVERSITY OF MARYLAND

$$e^* = \text{argmax}_e \, p(e|f;\theta)$$

Nature of data matters more in Neural MT

This Talk: Data Divergences in NMT

Examine implicit equivalence assumptions about bitext and MT

Show that divergences from these assumptions occur and matter for neural MT

Translation Divergences

"the same information is conveyed in the source and target text, but the structure of the sentences are different"

[Dorr 1994]

```
en: Maria did not slap
the green witch

es: Maria no daba una
botefada a la bruja verde
```

Divergence (according to WordNet)

- S: (n) **divergence**, divergency
(the act of moving away in different direction from a common point)

- S: (n) deviation, **divergence**, departure, difference
(a variation that deviates from the standard or norm)

Semantic Divergences

Assumption:

source and target side in bitext have the same meaning

Our hypothesis:

bitext sides are not always semantically equivalent and this matters for NMT

Reference Divergences

Assumption:

References can substitute for predicted translations during training

Our hypothesis:

Modeling divergences between references and predictions improves NMT

Style Divergences

Assumption:

MT output should preserve all properties of input

Our hypothesis:

We can tailor NMT style while preserving input meaning

Semantic Divergences

Reference Divergences

Style Divergences

Semantic Divergences

Assumption:

source and target side in bitext have the same meaning

Yet:

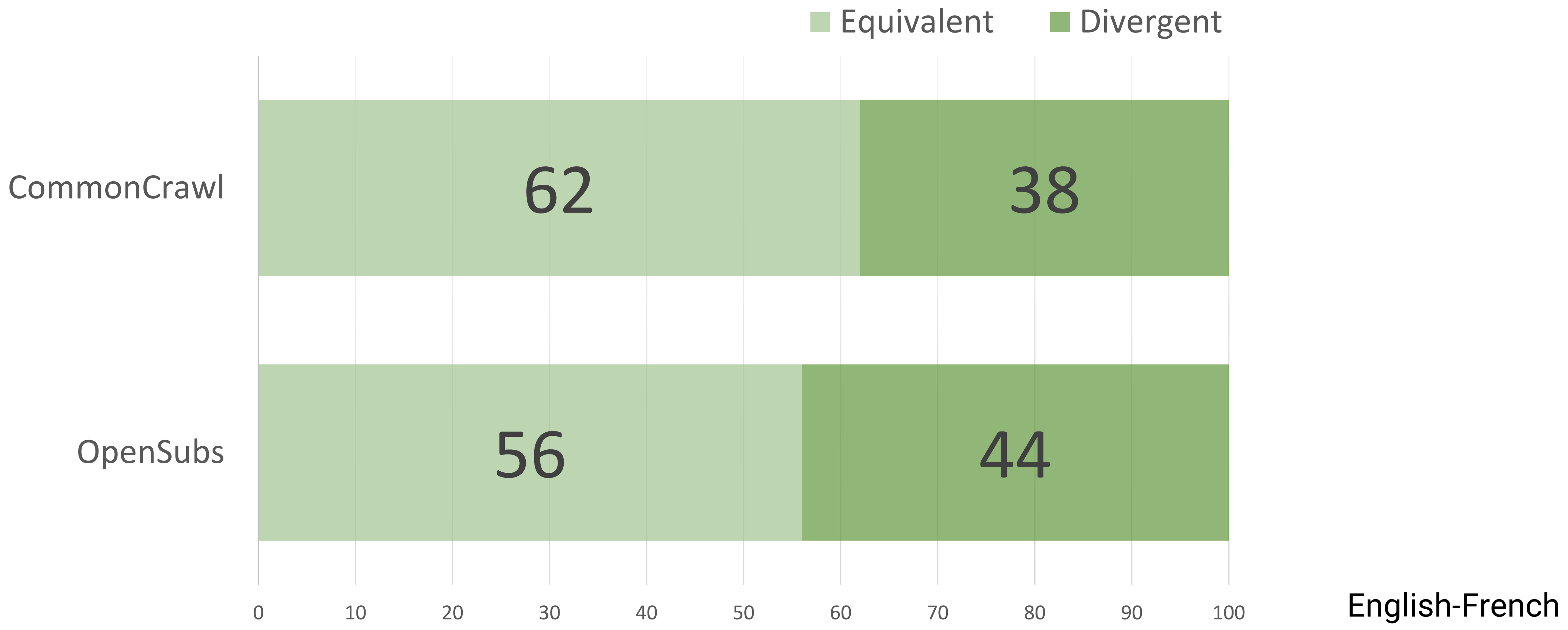
parallel documents ≠ parallel segments

“traduttore, traditore”: translators can alter source meaning

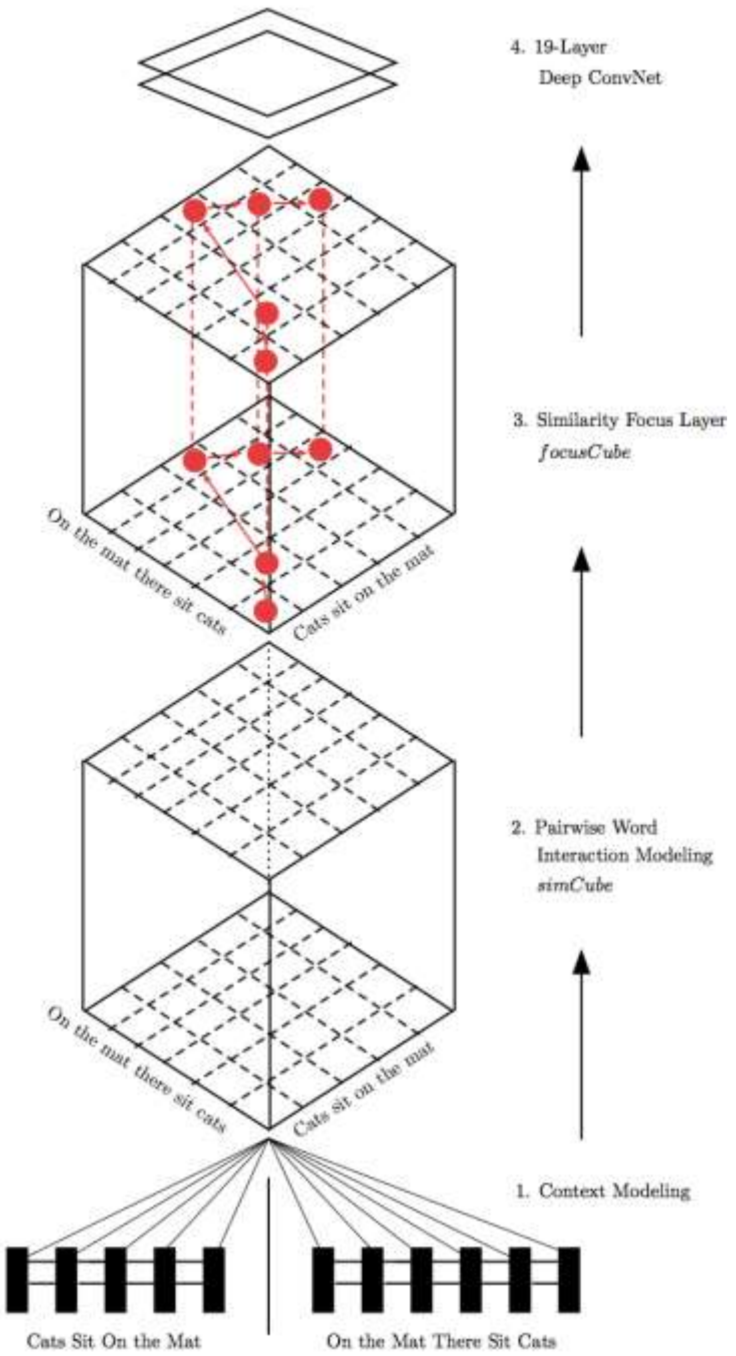
Divergence Examples

```
En: i don't know what i'm gonna do.
Fr: j'en sais rien.

En: you help me with zander and i helped you with joe.
Fr: tu m'as aidee avec zander, je t'ai aidee avec joe.

En: - has the sake chilled? - no, it's fine.
Fr: - c'est assez chaud?
```

How Frequent are Divergent Examples? A Crowdsourcing Experiment



Approach: cross-lingual semantic similarity model



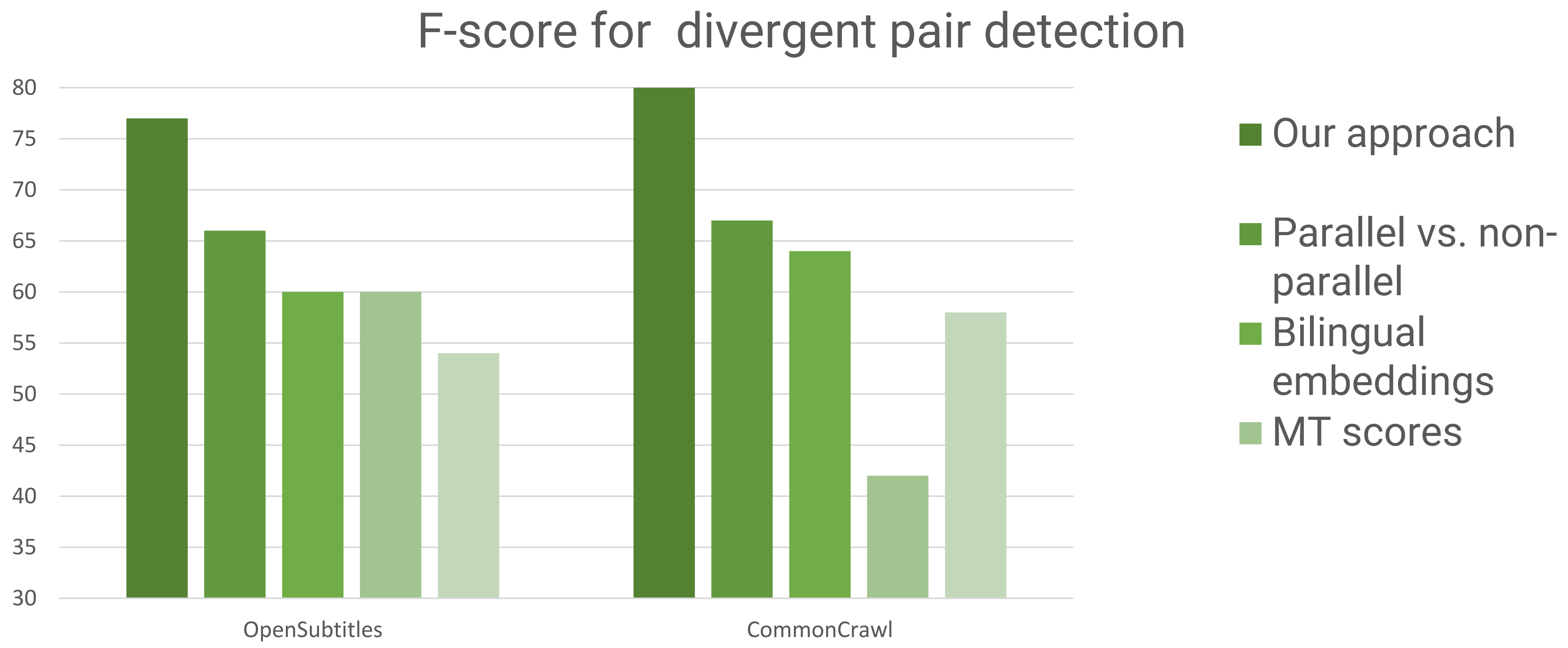
Predict semantic similarity with the "Very Deep Pairwise Similarity Model" [He & Lin 2016]

Initialize with bilingual word embeddings

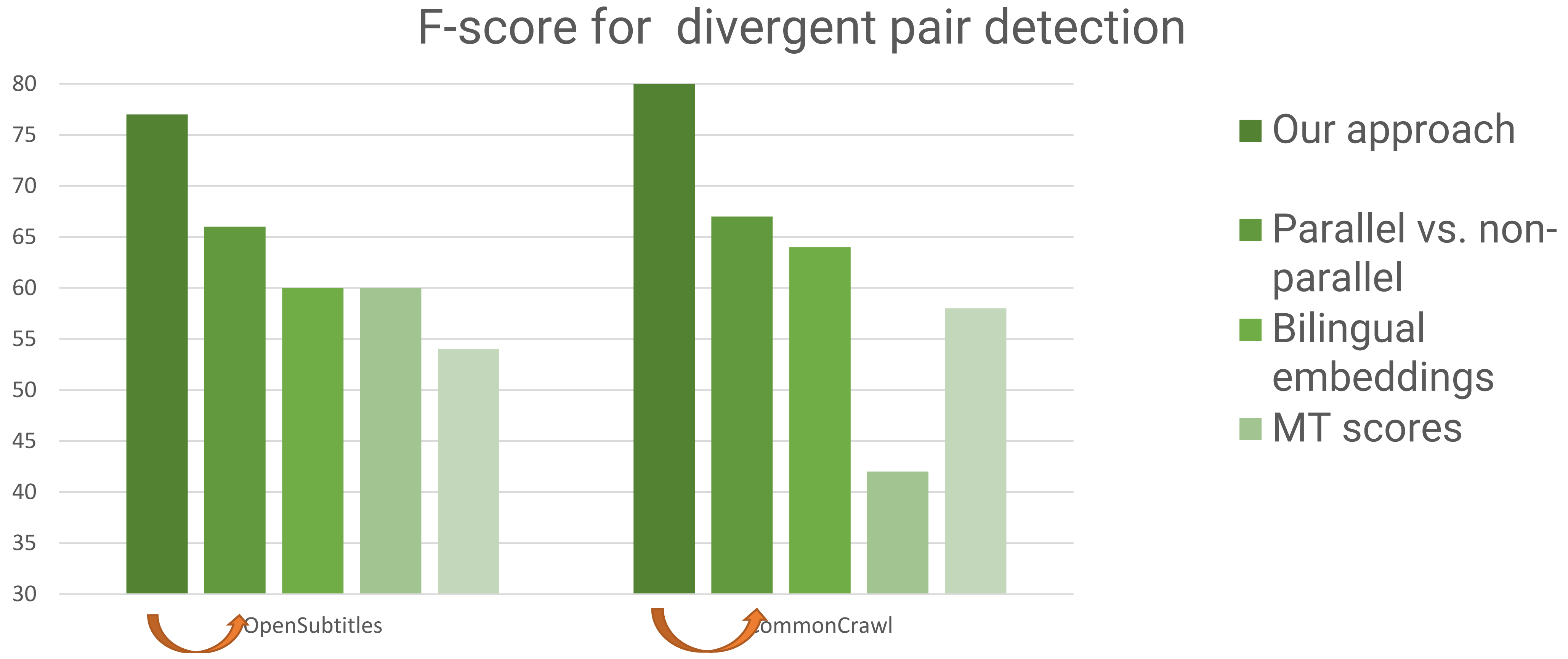
Approach: Generate (Noisy) Synthetic Training Examples

[Munteanu & Marcu 2006]

Intrinsic Evaluation: ConvNet trained on synthetic examples performs best

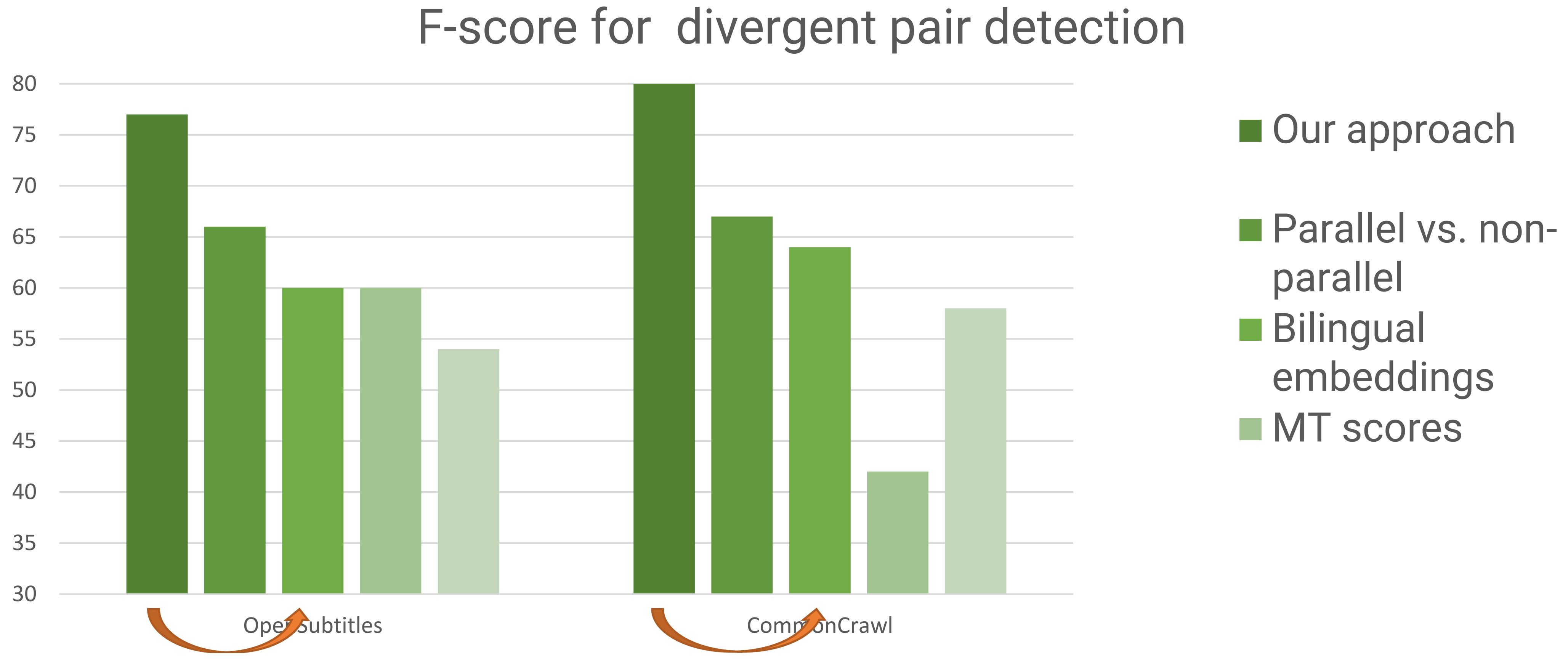


Intrinsic Evaluation: ConvNet trained on synthetic examples performs best



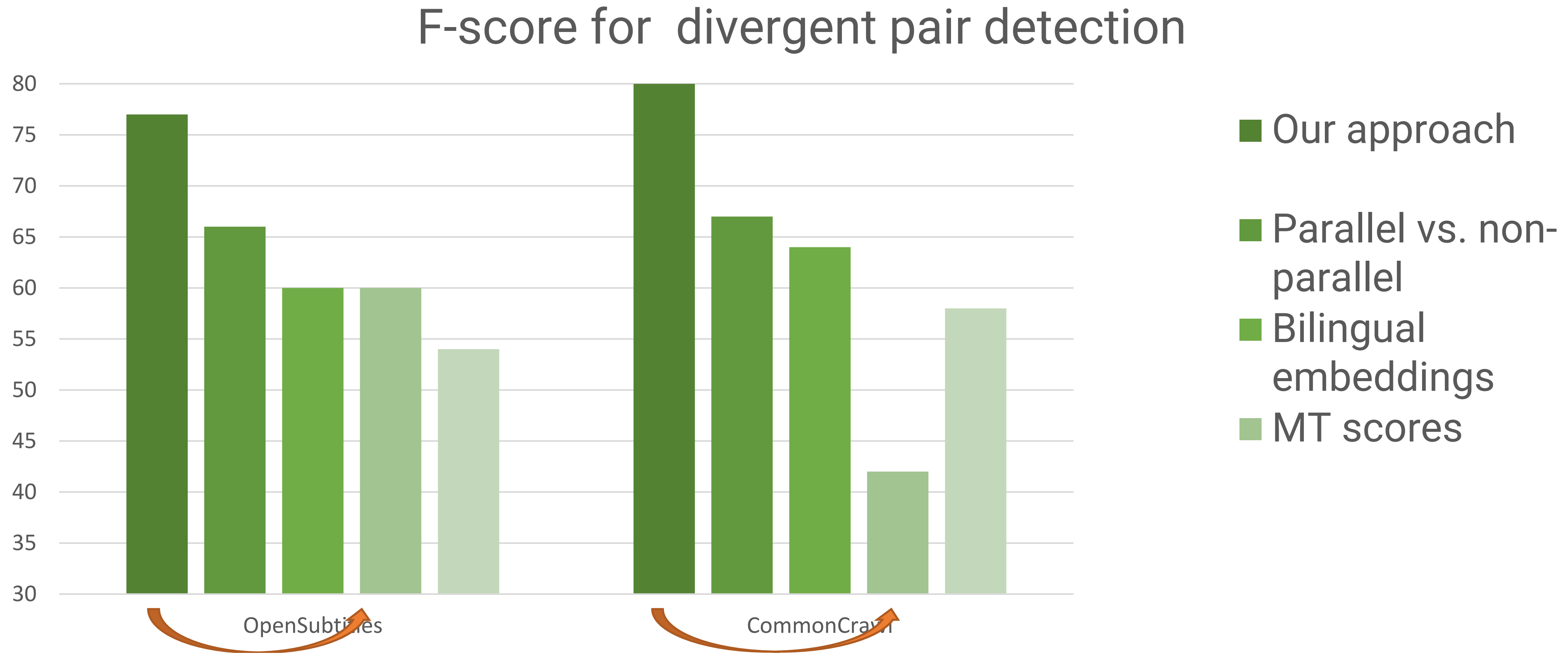
Worse F-score when using same synthetic examples with non-neural classifier [Munteanu & Marcu 2006]

Intrinsic Evaluation: ConvNet trained on synthetic examples performs best



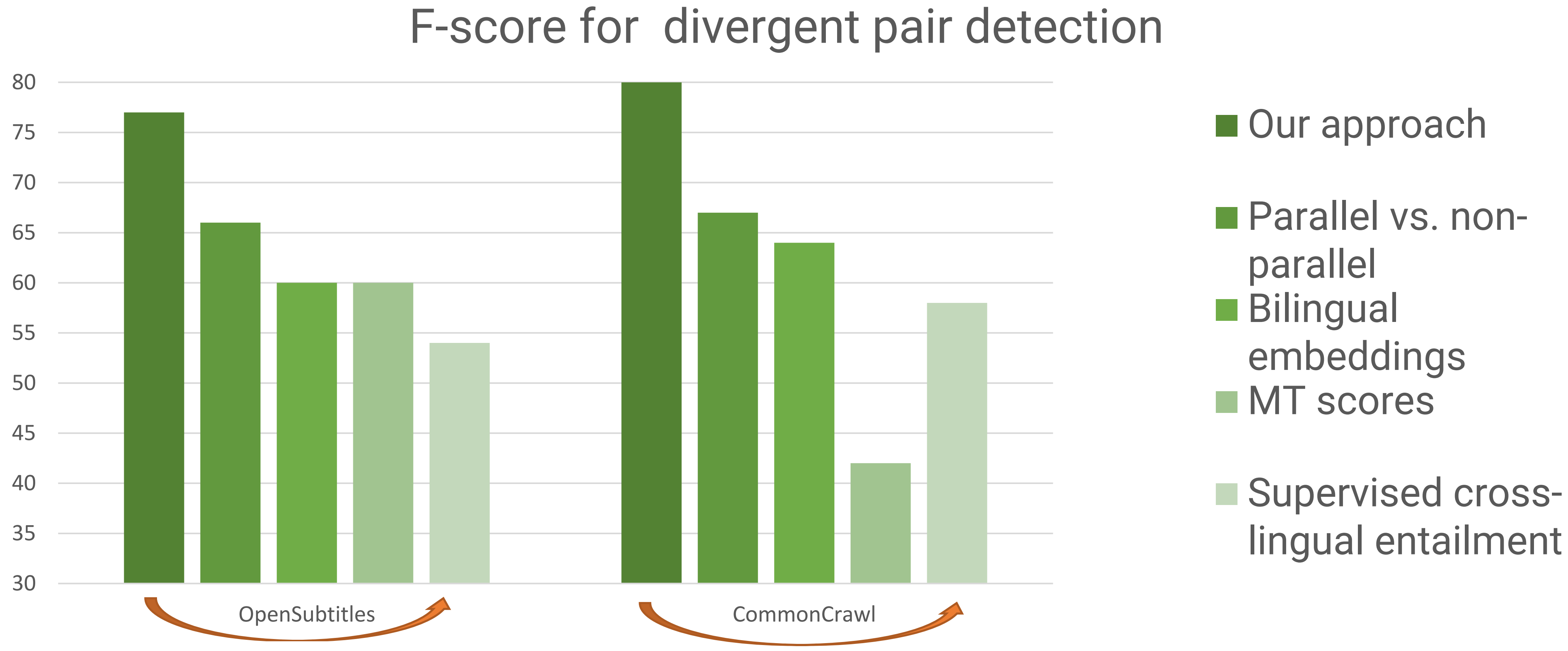
Worse F-score when using only bilingual word embeddings

Intrinsic Evaluation: ConvNet trained on synthetic examples performs best



Worse F-score when using NMT scores

Intrinsic Evaluation: ConvNet trained on synthetic examples performs best



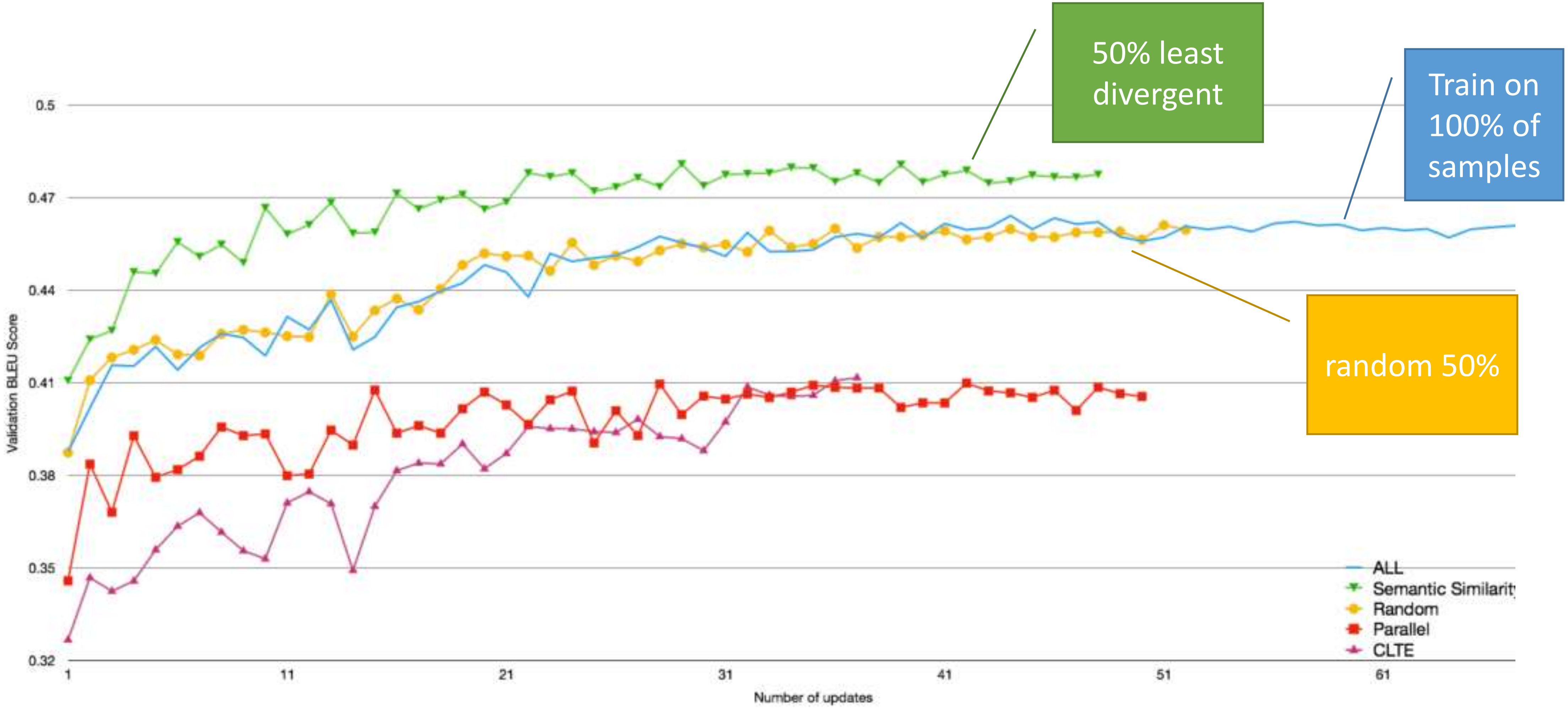
Worse F-score when using a supervised cross-lingual entailment classifier [Carpuat et al. 2017]

Do semantic divergences impact MT?

English > French tasks from IWSLT

Training Set	OpenSubtitles	33.5M segment pairs
In domain Test Set	MSLT: Microsoft Speech Language Translation (IWSLT16)	5000 segment pairs
Out of domain Test Set	TED talks (IWLST15)	1300 segment pairs

Downsampling via cross-lingual semantic similarity helps NMT training



[Vyas, Niu & Carpuat, NAACL 2018]

Downsampling via cross-lingual semantic similarity doesn’t hurt BLEU at test time

Model	MSLT BLEU		TED BLEU	
	Avg.	Ensemble	Avg.	Ensemble
RANDOM	43.49	45.64	36.05	38.20
PARALLEL	40.65	42.12	35.99	37.86
ENTAILMENT	39.64	41.86	33.30	35.40
SEMANTIC SIM.	**45.53**	**47.23***	**36.98**	**38.87**
ALL	44.64	46.26	36.98	38.59

[Vyas, Niu & Carpuat, NAACL 2018]

Beyond filtering divergent examples

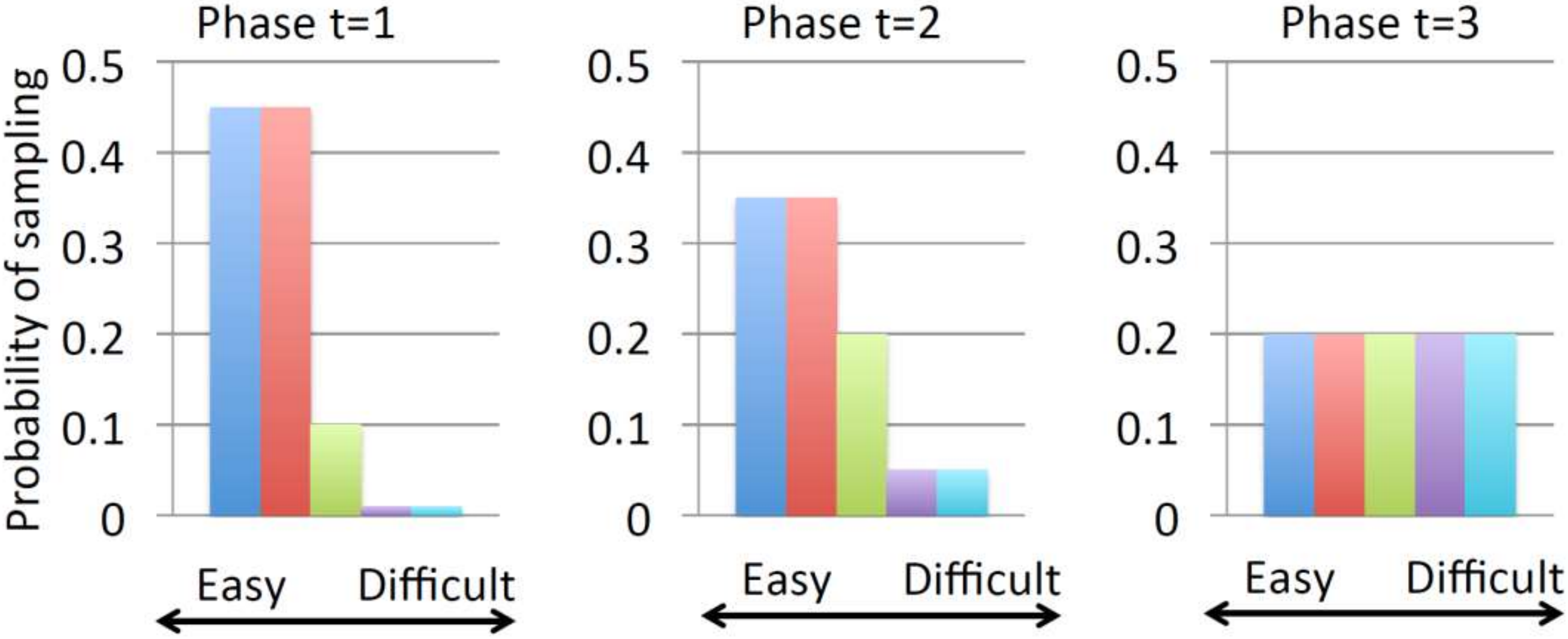
Fixing divergences by deleting extra info

[Pham et al. EMNLP 2018]

Curriculum learning with noise & domain criteria

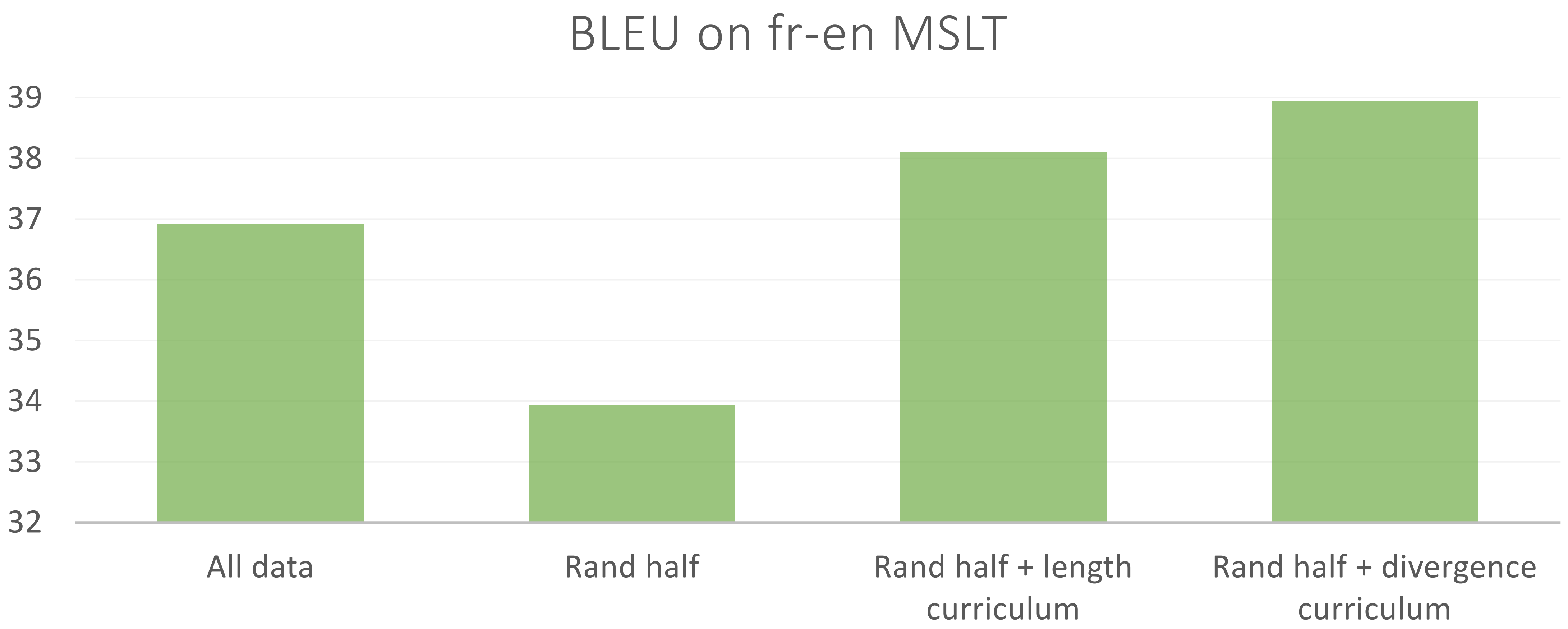
[Wang et al. NAACL 2019]

A Probabilistic Curriculum for Sampling Training Data



[Zhang et al. NAACL 2019]

Preview: Divergence-based Curriculum improves BLEU



[Richburg & Carpuat, unpublished]

Semantic Divergences

All bitexts contain semantically divergent examples

We can detect them with deep semantic similarity models trained on synthetic examples

Neural machine translation is sensitive to such divergences

Filtering out divergent examples helps

Open questions

What kind of divergences? How do they differ from noise?

Semantic Divergences

Curriculum Learning for Domain Adaptation in Neural Machine Translation. Xuan Zhang, Pamela Shapiro, Gaurav Kumar, Paul McNamee, Marine Carpuat and Kevin Duh. NAACL 2019

Identifying Semantic Divergences in Parallel Text without Annotations. Yogarshi Vyas, Xing Niu and Marine Carpuat. NAACL 2018

Detecting Cross-Lingual Semantic Divergence for Neural Machine Translation. Marine Carpuat, Yogarshi Vyas and Xing Niu. ACL Workshop on Neural Machine Translation 2017

github.com/yogarshi/SemDiverge

github.com/kevinduh/sockeye-recipes

Reference Divergences

aka Exposure Bias

Assumption:

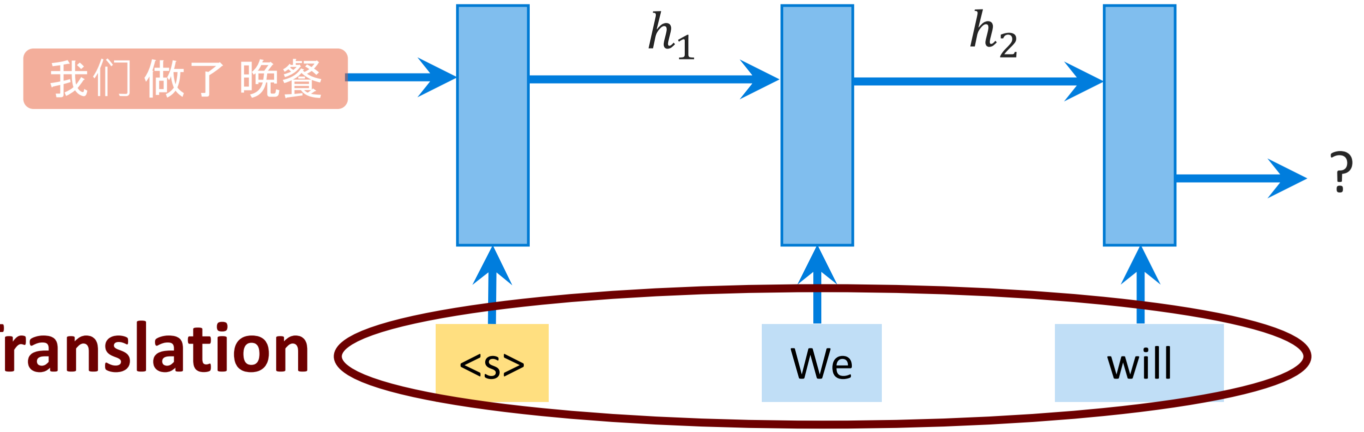
References can substitute for predicted translations during training

Our hypothesis:

Modeling divergences between references and predictions improves NMT

Exposure Bias: Gap Between Training and Inference

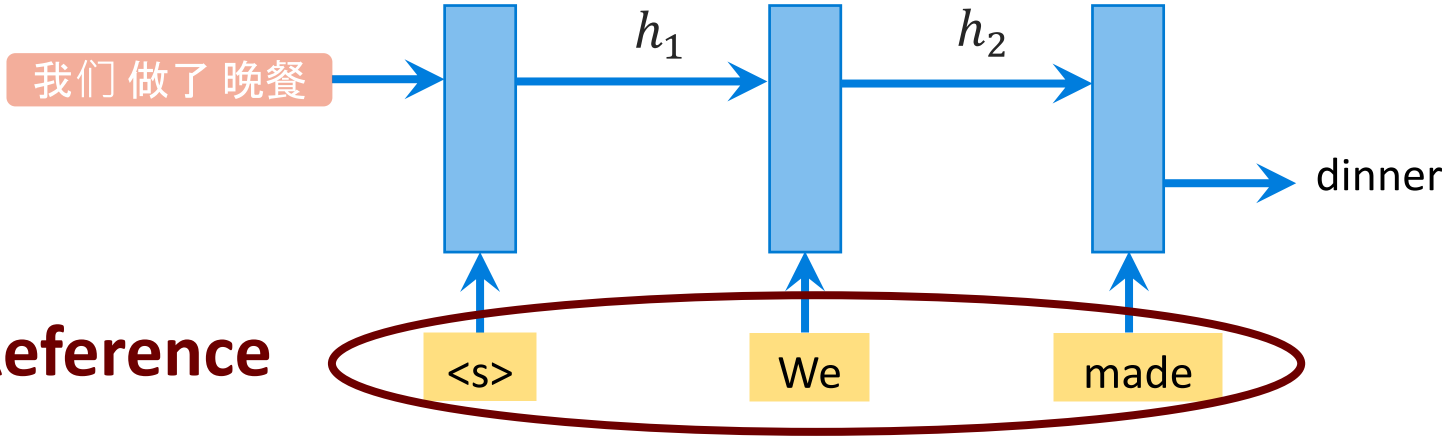
Inference



Model Translation

$$P(y|x) = \prod_{t=1}^{T} p(y_t | y_{<t}, x)$$

Maximum Likelihood Training



Reference

$$\text{Loss} = \sum_{t=1}^{T} \log p(y_t | y_{<t}, x)$$

How to Address Exposure Bias?

Expose models to their own predictions during training

But how to compute the loss when the partial translation diverges from the reference?

Our method: learn to align the reference words with partial translations during training.

Existing Methods

Search-based Methods

[Liang et al. 2006, Daumé et al. 2009, Leblond et al. 2017]

Computationally expensive

Reinforcement Learning with Sentence-Level Reward

[Ranzato et al., 2015, Bahdanau et al., 2016]

Inefficient and unstable

Scheduled Sampling

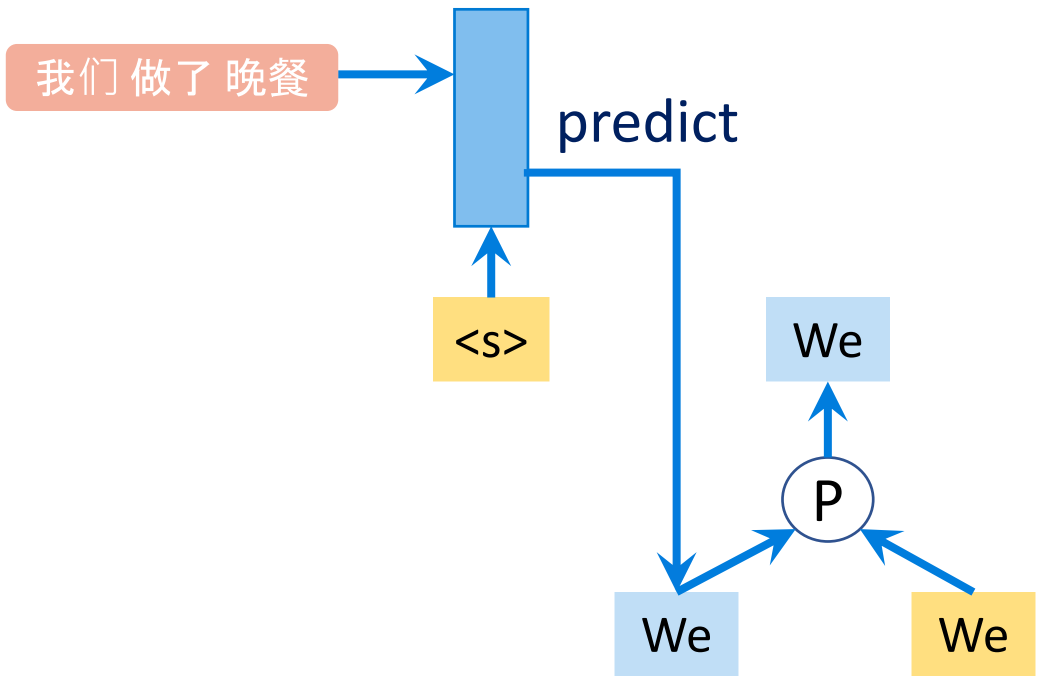
[Venkatraman et al. 2015, Bengio et al. 2015, Goyal et al. 2017]

Simple and efficient, but ...

Existing Method: Scheduled Sampling

Reference: <s> We made dinner </s>

P = choose randomly

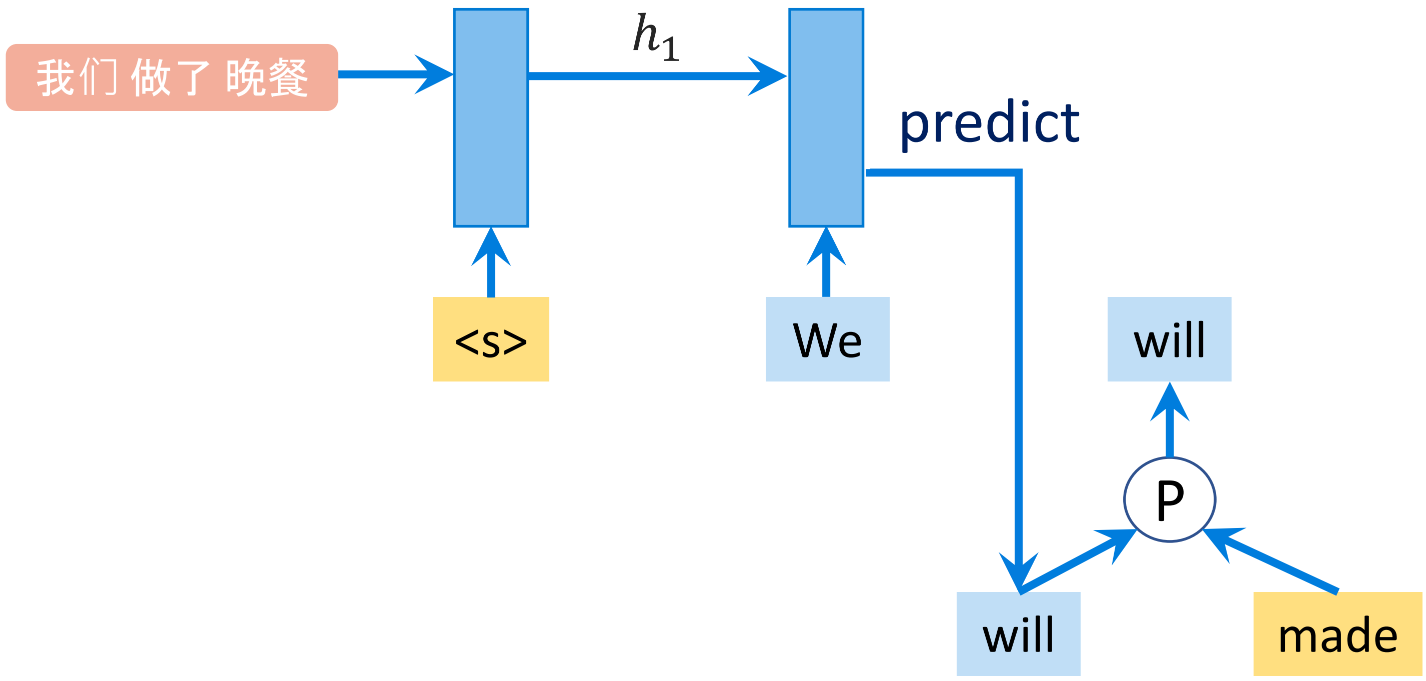


[Bengio et al., NeurIPS 2015]

Existing Method: Scheduled Sampling

Reference: <s> We made dinner </s>

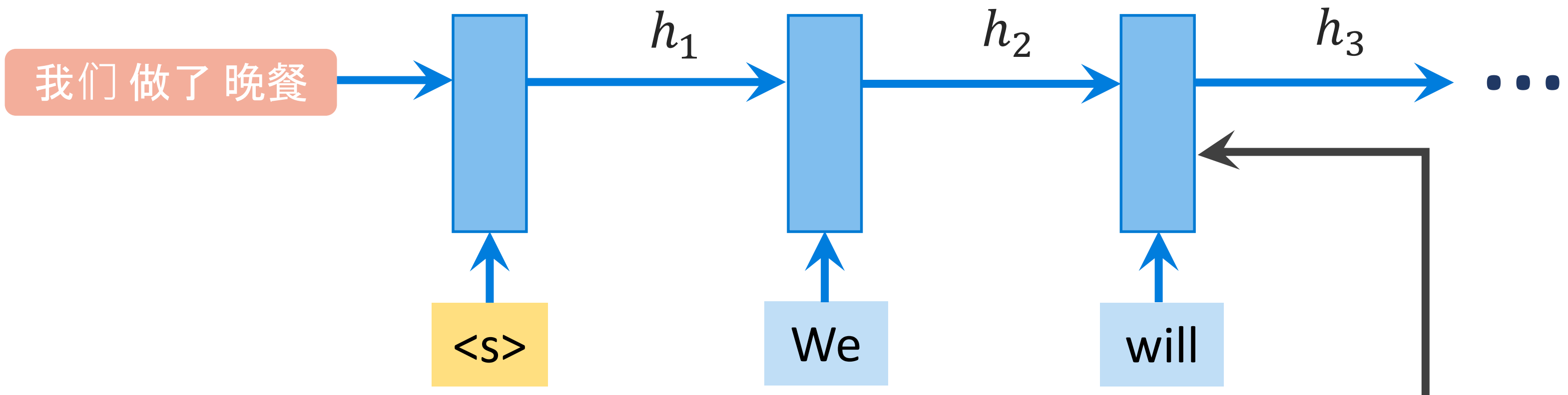
P = choose randomly



[Bengio et al., NeurIPS 2015]

Existing Method: Scheduled Sampling

Reference: <s> We made dinner </s>

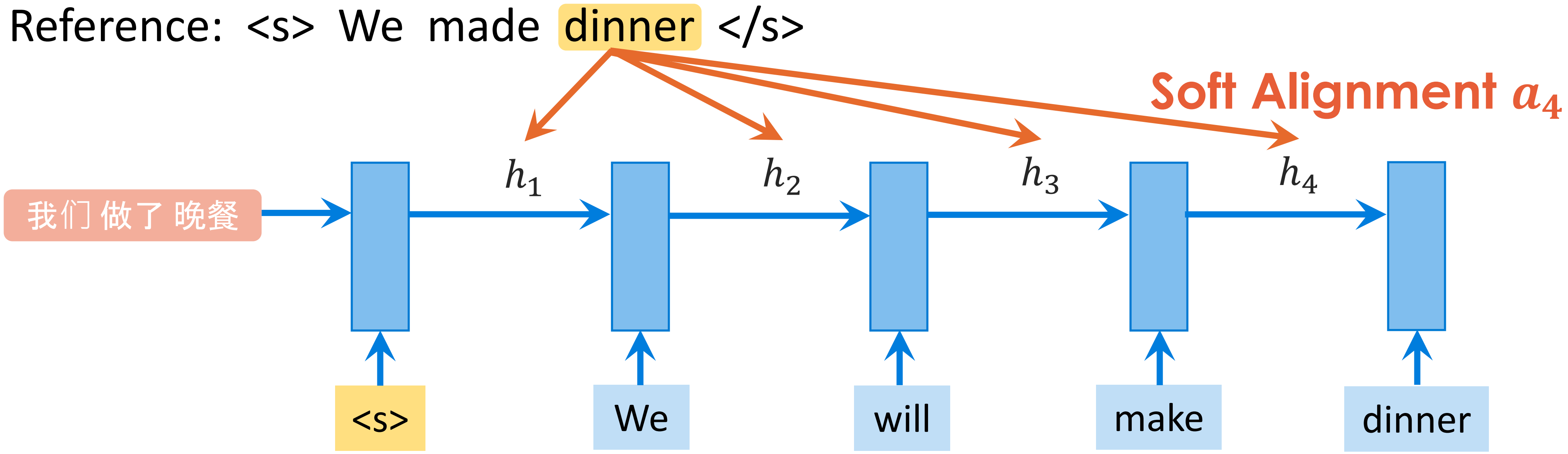


Incorrect synthetic reference:
“We will dinner”

J = log p(“dinner” | “<s> We will”, source)

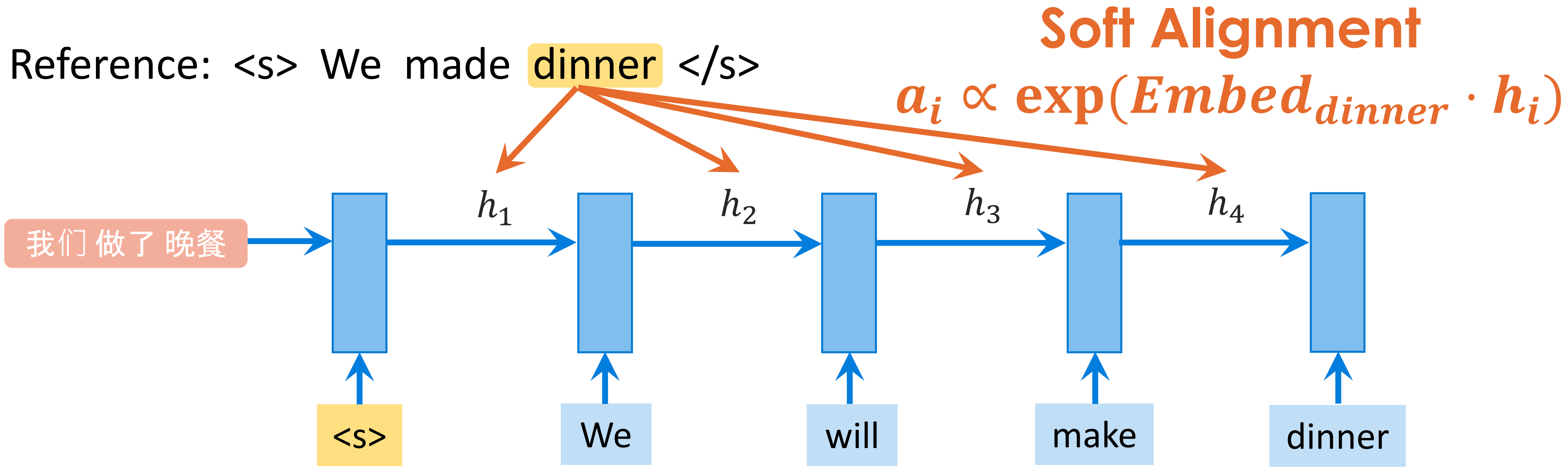
[Bengio et al., NeurIPS 2015]

Our Solution: Learning How To Align Reference with Partial Translations



a_1 logp(“dinner” | “<s>”, source) + a_2 logp(“dinner” | “<s> We”, source) + a_3 logp(“dinner” | “<s> We will”, source) + a_4 logp(“dinner” | “<s> We will make”, source)

Our Solution: Learning How To Align Reference with Partial Translations

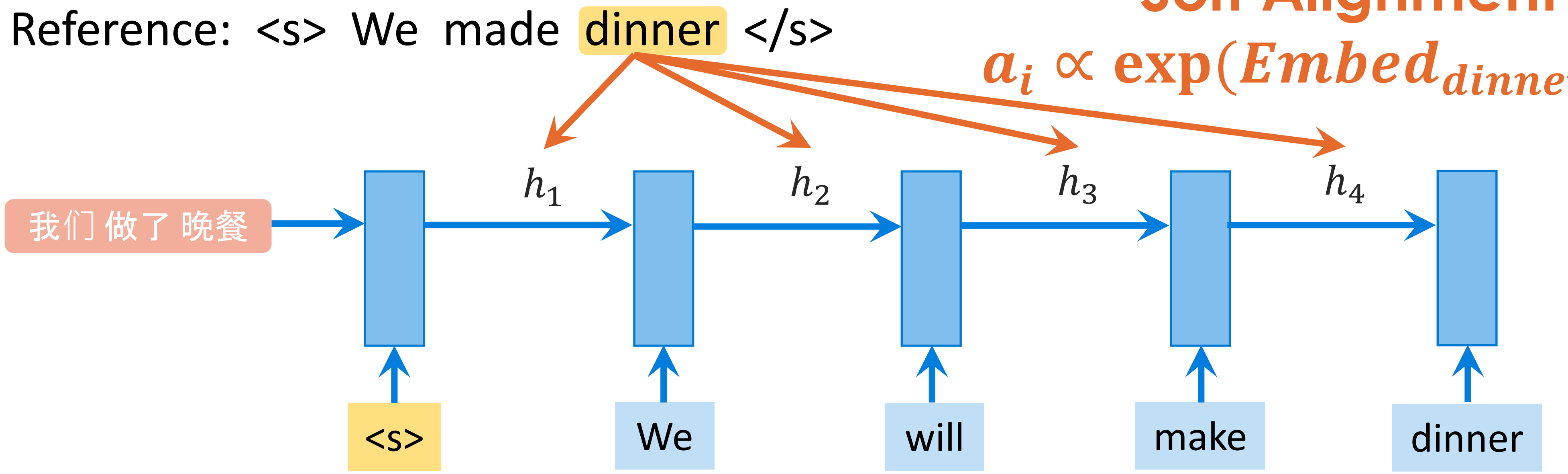


a_1 logp(“dinner” | “<s>”, source) + a_2 logp(“dinner” | “<s> We”, source) +
a_3 logp(“dinner” | “<s> We will”, source) + a_4 logp(“dinner” | “<s> We will make”, source)

Our Solution: Learning How To Align Reference with Partial Translations

Soft Alignment

$$\boldsymbol{a_i \propto \exp(Embed_{dinner} \cdot h_i)}$$



$\boldsymbol{a_1}$ logp("dinner" | "<s>", source) + $\boldsymbol{a_2}$ logp("dinner" | "<s> We", source) + $\boldsymbol{a_3}$ logp("dinner" | "<s> We will", source) + $\boldsymbol{a_4}$ **logp("dinner" | "<s> We will make", source)**

Training Objective

Ours:

Soft alignment between y_t and $\tilde{y}_{<j}$

$$J_{SA} = \sum_{(x,y)\in D} \sum_{t=1}^{T} log \sum_{j=1}^{T'} a_{tj}\, p(y_t \mid \tilde{y}_{<j}, x)$$

Scheduled Sampling:

Hard alignment by time index t

$$J_{SS} = \sum_{(x,y)\in D} \sum_{t=1}^{T} log\, p(y_t \mid \tilde{y}_{<t}, x)$$

Training Objective

Ours:

Soft alignment between y_t and $\tilde{y}_{<j}$

$$J_{SA} = \sum_{(x,y)\in D} \sum_{t=1}^{T} log \sum_{j=1}^{T'} a_{tj}\, p(y_t \mid \tilde{y}_{<j}, x)$$

Scheduled Sampling:

Hard alignment by time index t

$$J_{SS} = \sum_{(x,y)\in D} \sum_{t=1}^{T} log\, p(y_t \mid \tilde{y}_{<t}, x)$$

Training Objective

Ours:

Soft alignment between y_t and $\tilde{y}_{<j}$

$$J_{SA} = \sum_{(x,y)\in D} \sum_{t=1}^{T} log \sum_{j=1}^{T'} a_{tj}\, p(y_t \mid \tilde{y}_{<j}, x)$$

Scheduled Sampling:

Hard alignment by time index t

$$J_{SS} = \sum_{(x,y)\in D} \sum_{t=1}^{T} log\, p(y_t \mid \tilde{y}_{<t}, x)$$

Combined with maximum likelihood:

$$J = J_{SA} + J_{ML}$$

Experiments

Data

IWSLT14 de-en

IWSLT15 vi-en

Task	sentences (K)			vocab (K)	
	train	dev	test	src	tgt
de-en	153.3	7.0	6.8	113.5	53.3
vi-en	121.3	1.5	1.3	23.9	50.0

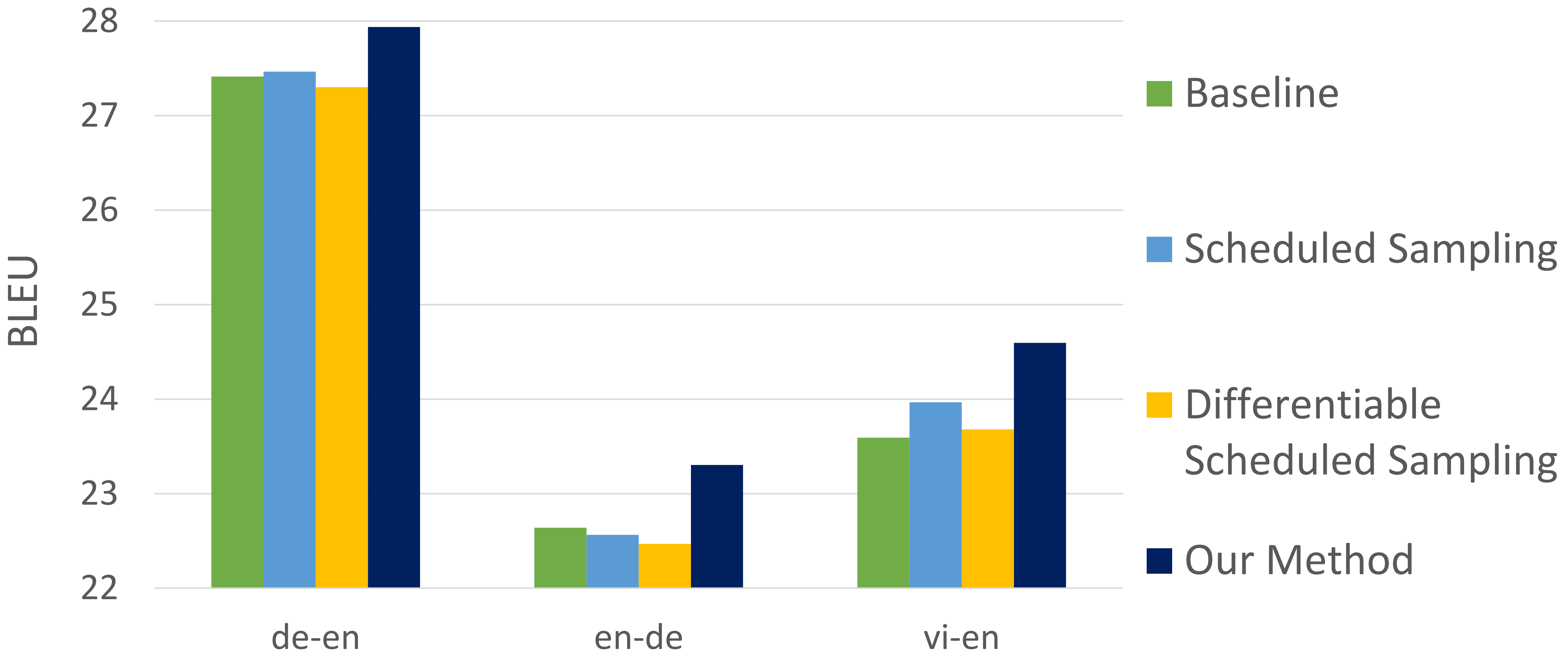
Model

Bi-LSTM encoder, LSTM decoder, multilayer perceptron attention

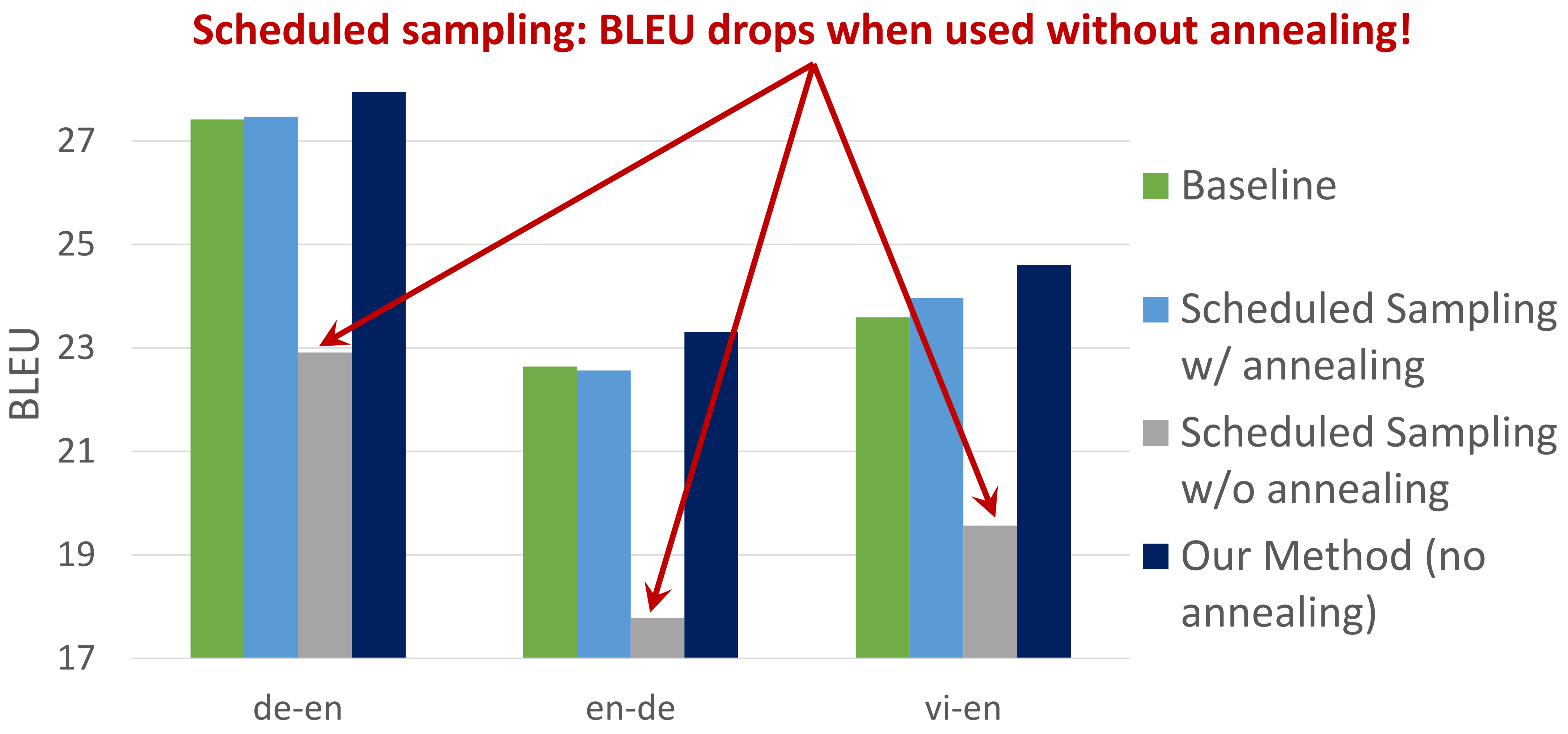
Differentiable sampling with Straight-Through Gumbel Softmax

Based on AWS sockeye

Our Method Outperforms Maximum Likelihood and Scheduled Sampling



Our Method Needs No Annealing



Reference Divergences

A new training objective

1. **Generate translation prefixes** via differentiable sampling
2. Learn to **align** the reference words with sampled prefixes

Better BLEU than the maximum likelihood and scheduled sampling (de-en, en-de, vi-en)

Simple to train, no annealing schedule required

Reference Divergences

Flexible Reference Word Order for Neural Machine Translation

Weijia Xu, Xing Niu, Marine Carpuat. NAACL 2019

github.com/lzecson/saml-nmt

Style Divergences

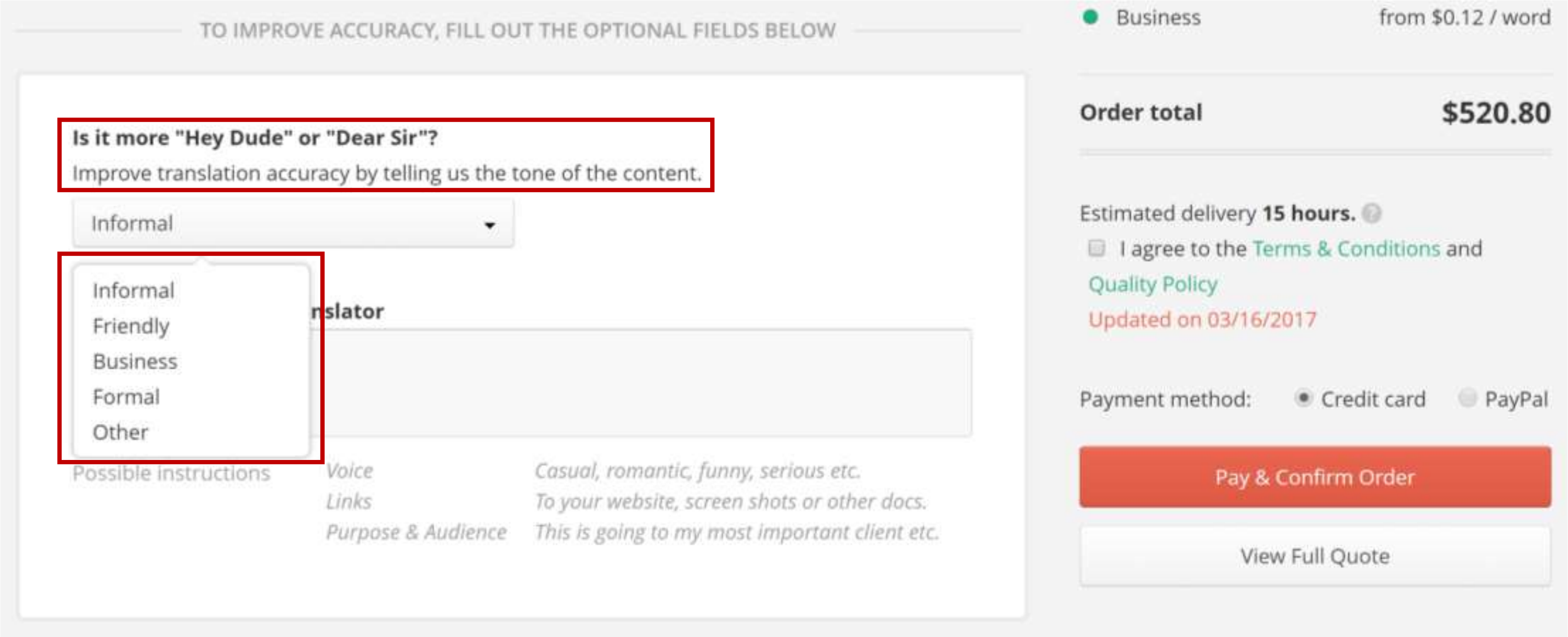
Assumption:

MT output should preserve all properties of input

Our hypothesis:

We can tailor NMT style while preserving input meaning

Style Matters for Translation



www.gengo.com

Does Style Matter for Machine Translation?

We focus on **formality**

Goal: Can we produce MT output with varying formality?

Prior work: other aspects of style

- conversational language [Lewis et al. 2015]
- politeness (du vs. Sie) [Sennrich et al. 2016]
- personalization (gender) [Rabinovich et al. 2017]

Formality-Sensitive Machine Translation (FSMT)

Comment ça va? Source (f) → FSMT (θ) → Translation-1 (e_1) How are you doing?

or

Desired formality level (ℓ) → Translation-2 (e_2) What's up?

How to train?

f	ℓ_1	e_1
f	ℓ_2	e_2

Ideal training data doesn't occur naturally!

[Niu, Martindale & Carpuat, EMNLP 2017]

Formality in MT Corpora

Formal

Informal

`delegates are kindly requested to bring their copies of documents to meetings .`	[UN]
`in these centers , the children were fed , medically treated and rehabilitated on both a physical and mental level .`	[OpenSubs]
`there can be no turning back the clock`	[UN]
`I just wanted to introduce myself`	[OpenSubs]
`-yeah , bro , up top .`	[OpenSubs]

Formality Transfer (FT)

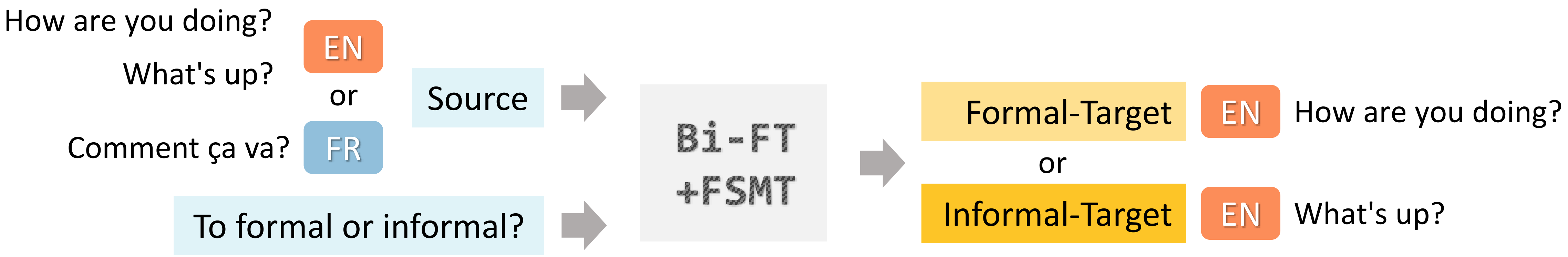
What's up? EN Informal-Source → FT → Formal-Target EN How are you doing?

How are you doing? EN Formal-Source → FT → Informal-Target EN What's up?

Given a large parallel formal-informal corpus
(e.g., Grammarly's Yahoo Answers Formality Corpus)
these are sequence-to-sequence tasks

[Rao and Tetreault, 2018]

Formality Sensitive MT as Multitask Formality Transfer + MT



Multitask Formality Transfer + MT

Model: shared encoder, shared decoder as in multilingual NMT [Johnson et al. 2017]

Training objective:

$$\mathcal{L}_{MT} + \mathcal{L}_{FT}$$

$$\mathcal{L}_{MT} = \sum_{(\boldsymbol{X},\boldsymbol{Y})} \log P(\boldsymbol{Y}|\boldsymbol{X};\boldsymbol{\theta})$$

MT pairs

$$\mathcal{L}_{FT} = \sum_{(\mathbf{Y}_{\bar{\ell}},\mathbf{Y}_{\ell})} \log P(\mathbf{Y}_{\ell} \,|\, \mathbf{Y}_{\bar{\ell}}, \ell; \boldsymbol{\theta})$$

FT pairs

Multitask Formality Transfer + MT Training Data

FT	<F>	Informal-EN	Formal-EN
	<I>	Formal-EN	Informal-EN

Side constraint
[Sennrich et al. 2016]

50k sentence pairs from Grammarly's Yahoo Answers Formality Corpus

Multitask Formality Transfer + MT Training Data

FT	<F>	Informal-EN	Formal-EN
	<I>	Formal-EN	Informal-EN
MT	<F>	FR	Formal-EN
	<I>	FR	Informal-EN

Data selected
[Moore & Lewis, 2010]
from OpenSubtitles

Evaluation – Formality Transfer

Test set

Grammarly's Yahoo
Answers Formality Corpus

1K sent pairs per direction
4 references
Automatic metric: BLEU

[Rao & Tetreault, 2018]

Multitask Model

Model

1 layer LSTM encoder decoder
MLP attention

Shared 30k BPE vocab
Tied src emb, trg emb, output layer
512 embeddings, hidden layers

Toolkit: AWS Sockeye

Results – Formality Transfer (BLEU)

Model	Informal→Formal E&M	Informal→Formal F&R	Formal→Informal E&M	Formal→Informal F&R
Original Source	49.09	51.03	29.85	29.85
PBMT (Rao and Tetreault, 2018)	68.22	72.94	33.54	32.64
NMT Baseline (Rao and Tetreault, 2018)	58.80	68.28	30.57	36.71
NMT Combined (Rao and Tetreault, 2018)	68.41	74.22	33.56	35.03
NMT Baseline	65.34	71.28	32.36	36.23
Bi-directional FT	66.30	71.97	34.00	36.33
+ training on E&M + F&R	69.20	73.52	35.44	37.72
+ ensemble decoding (×4)	71.36	74.49	36.18	38.34
+ multi-task learning	**72.13**	**75.37**	**38.04**	**39.09**

Results – Formality Transfer (BLEU)

Model	Informal→Formal E&M	F&R	Formal→Informal E&M	F&R
Original Source	49.09	51.03	29.85	29.85
PBMT (Rao and Tetreault, 2018)	68.22	72.94	33.54	32.64
NMT Baseline (Rao and Tetreault, 2018)	58.80	68.28	30.57	36.71
NMT Combined (Rao and Tetreault, 2018)	68.41	74.22	33.56	35.03
NMT Baseline	65.34	71.28	32.36	36.23
Bi-directional FT	66.30	71.97	34.00	36.33
+ training on E&M + F&R	69.20	73.52	35.44	37.72
+ ensemble decoding (×4)	71.36	74.49	36.18	38.34
+ multi-task learning	**72.13**	**75.37**	**38.04**	**39.09**

Results – Formality Transfer Human Evaluation

Model	**Formality Difference I-F** Range = [0,2]	**Formality Difference F-I** Range = [0,2]	**Meaning Preservation** Range = [0,3]
Rao&Tetreault baseline	0.54	0.45	2.94
Multitask FT+MT	0.59	**0.64**	2.92

300 samples per model
3 judgments per sample
Protocol based on Rao & Tetreault

Multitask Formality Transfer + MT Training Data

FT	<F>	Informal-EN	Formal-EN
	<I>	Formal-EN	Informal-EN
MT	<F>	FR	Formal-EN
	<I>	FR	Informal-EN

Data selected [Moore & Lewis, 2010] from OpenSubtitles

Selected bilingual data is similar to GYAFC (FT☺)
GYAFC ≠ domain of translation data (FSMT☹)

Multitask Formality Transfer + MT Training Data Variants

MultiTask Select

<F>	Informal-EN	Formal-EN
<I>	Formal-EN	Informal-EN
<F>	FR	Formal-EN
<I>	FR	Informal-EN

<F>	FR	Formal-EN
<I>	FR	Informal-EN

Side constraint

MultiTask Rand

<F>	Informal-EN	Formal-EN
<I>	Formal-EN	Informal-EN
	FR	EN

Evaluation – Formality Sensitive MT

French-English

Training Data

- 50K pairs from GYAFC
- 2.5M pairs selected from OpenSubtitles 2016

Test

- Microsoft Spoken Language Corpus
- 1 reference of unknown formality

Formality Sensitive MT
BLEU ~~Evaluation~~

Model	FR to formal EN	FR to informal EN
MultiTask Select	25.02	25.20
MultiTask Rand	25.24	25.14
Side constraint	27.15	26.70
Phrase-based MT + formality reranking [Niu & Carpuat 2017]	29.12	29.02

Formality Transfer MT Human Evaluation

Model	**Formality Difference** Range = [0,2]	**Meaning Preservation** Range = [0,3]
MultiTask Rand	0.35	2.95
Side constraint	0.32	2.90
Phrase-based MT + formality reranking [Niu & Carpuat 2017]	0.05	2.97

300 samples per model
3 judgments per sample
Protocol based on Rao & Tetreault

Analysis: Multitask model makes more formality changes

	Reference	Refrain from the commentary and respond to the question, Chief Toohey.
Formal	MultiTask	**You need to be quiet** and answer the question, Chief Toohey.
	Side constraint	Please refrain from **any** comment and answer the question, Chief Toohey.
	PBMT	Please refrain from comment and **just** answer **the** question, **the** Tooheys's boss.
Informal	MultiTask	**Shut up** and answer the question, Chief Toohey.
	Side constraint	Please refrain from comment and answer the question, chief Toohey.
	PBMT	Please refrain from comment and answer **my** question, Tooheys's boss.

Analysis: Multitask model introduces more meaning errors

	Reference	Try to file any additional motions as soon as you can.
Formal	MultiTask	You should try to introduce the **sharks** as soon as you can.
	Side constraint	Try to present additional requests as soon as you can.
	PBMT	Try to introduce any additional requests as soon as you can.
Informal	MultiTask	Try to introduce **sharks** as soon as you can.
	Side constraint	Try to introduce extra requests as soon as you can.
	PBMT	Try to introduce any additional requests as soon as you can.

Preview: Improving Multitask Training with Synthetic Supervision

Multi Task Loss so far:

$$\mathcal{L}_{MT} + \mathcal{L}_{FT}$$

$$\mathcal{L}_{MT} = \sum_{\boxed{(\boldsymbol{X},\boldsymbol{Y})\text{ MT pairs}}} \log P(\boldsymbol{Y}|\boldsymbol{X};\boldsymbol{\theta})$$

$$\mathcal{L}_{FT} = \sum_{\boxed{(\boldsymbol{Y}_{\bar{\ell}},\boldsymbol{Y}_{\ell})\text{ FT pairs}}} \log P(\boldsymbol{Y}_{\ell}\,|\,\boldsymbol{Y}_{\bar{\ell}},\ell;\boldsymbol{\theta})$$

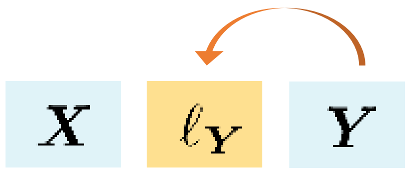
Hypothesis:

Training with complete FSMT examples can improve formality control while preserving meaning

$\boxed{(\boldsymbol{X},\ell,\boldsymbol{Y}_{\ell})\text{ FSMT triplets}}$

Improving Multitask Training with Synthetic Supervision

1. Online Style Inference (OSI): predict formality of MT samples on the fly



2. Replace MT loss by OSI loss

$$\mathcal{L}_{OSI} = \sum_{(\boldsymbol{X}, \ell_{\boldsymbol{Y}}, \boldsymbol{Y})} \log P(\boldsymbol{Y} \mid \boldsymbol{X}, \ell_{\boldsymbol{Y}}; \boldsymbol{\theta})$$

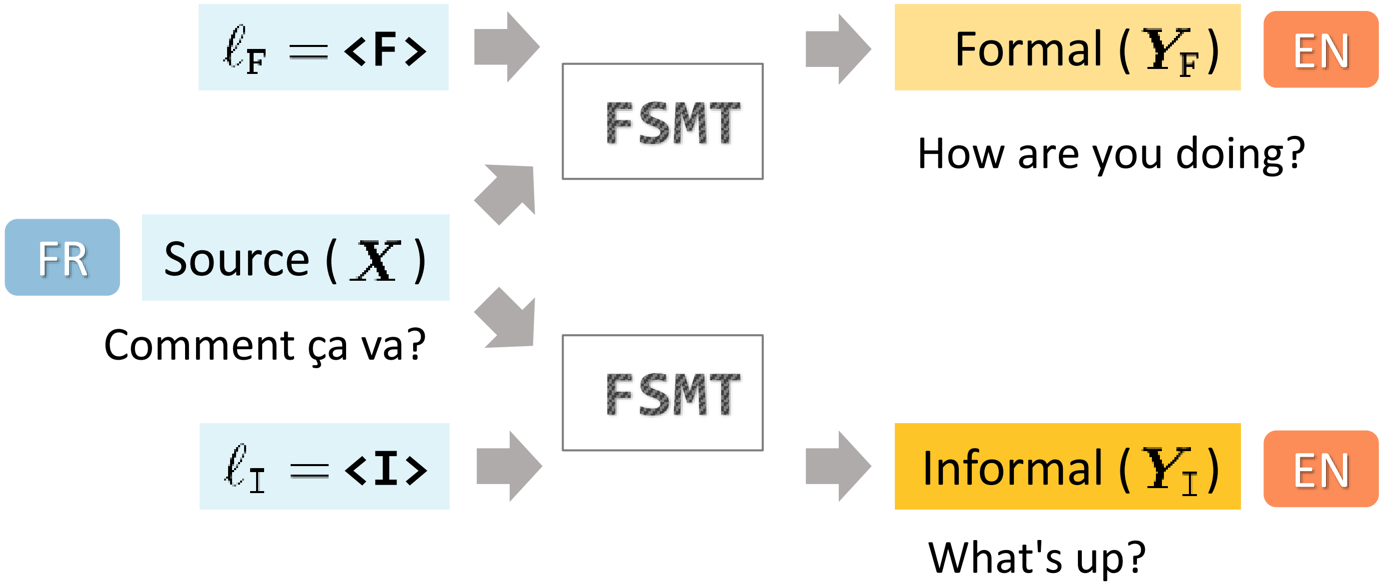
$$\mathcal{L} = \mathcal{L}_{FT} + \mathcal{L}_{OSI}$$

Synthetic Supervision: Predict formality of MT samples on the fly

By comparing reference to formal vs. informal translations of source

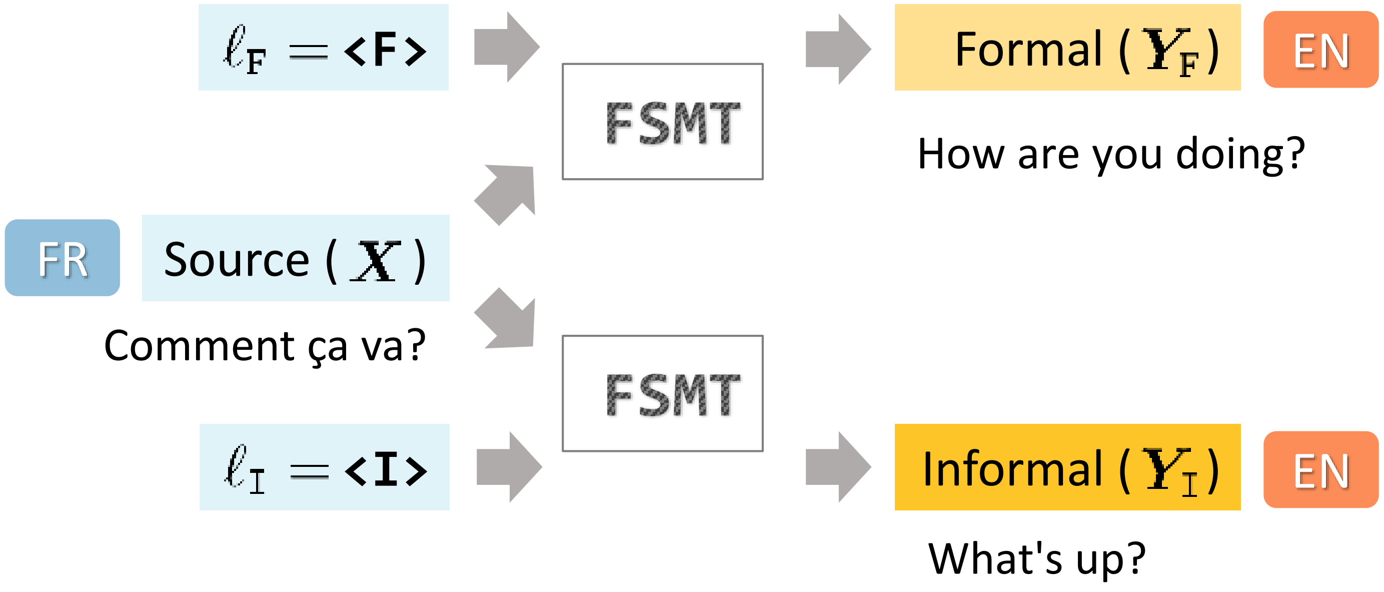
Synthetic Supervision: Predict formality of MT samples on the fly

By comparing reference to formal vs. informal translations of source



Synthetic Supervision: Predict formality of MT samples on the fly

By comparing reference to formal vs. informal translations of source

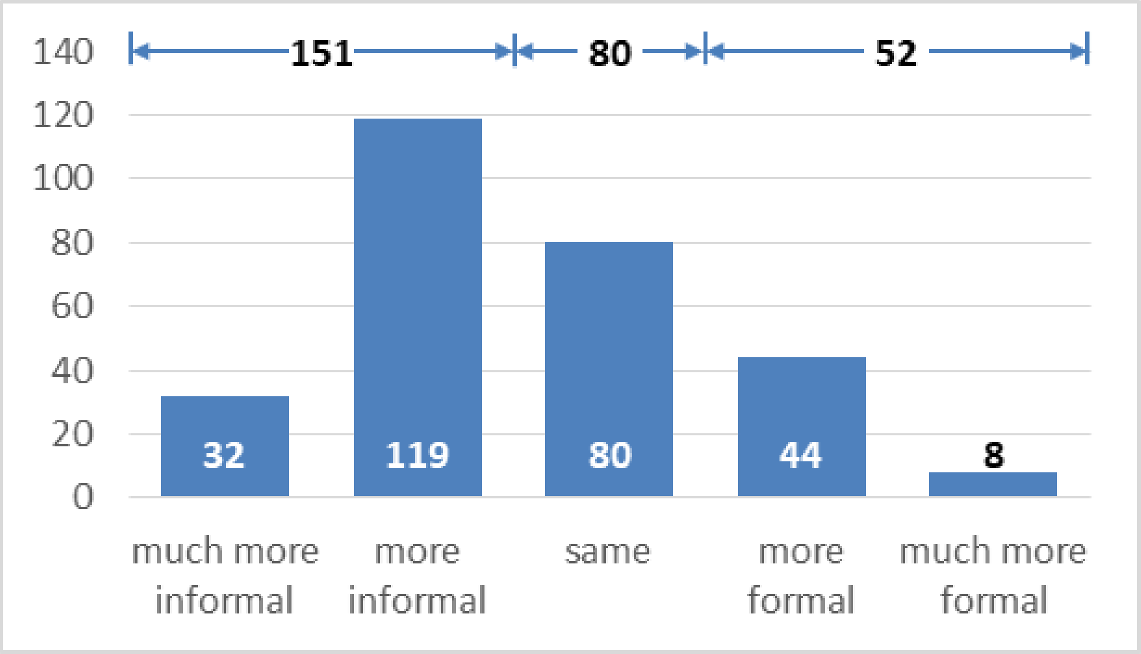


Target ($\boldsymbol{Y}$) EN How are you?

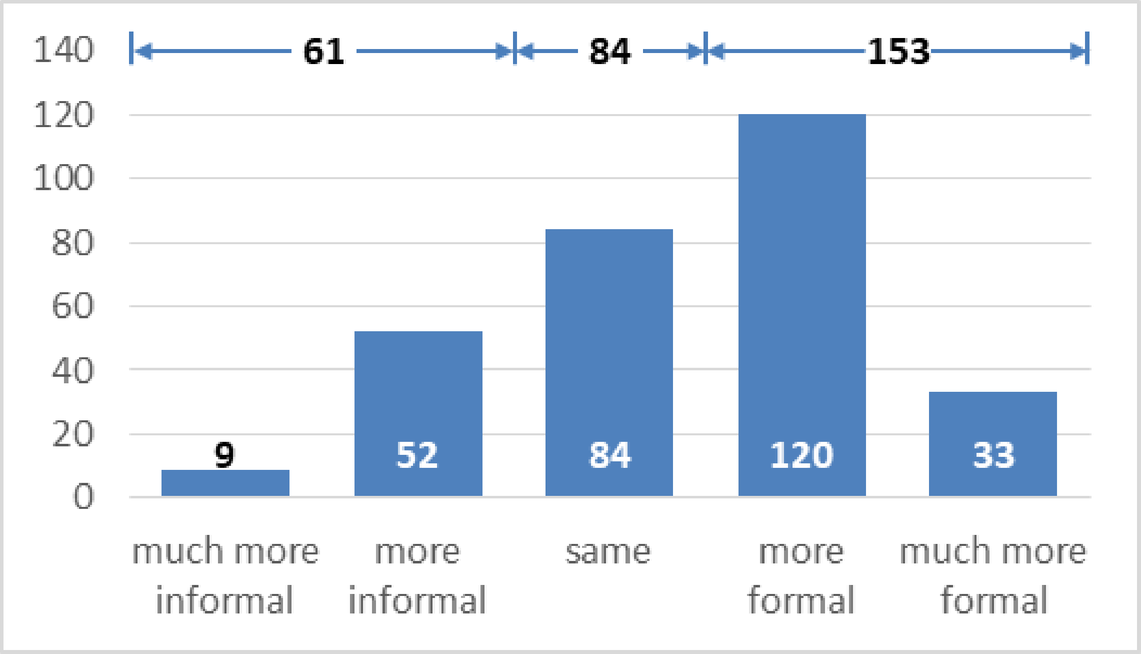
$$(\boldsymbol{X}, \boldsymbol{Y}) \Rightarrow (\boldsymbol{X}, \ell_{\mathrm{F}}, \boldsymbol{Y}) \text{ if } \mathrm{CED}(\boldsymbol{Y}_{\mathrm{I}}, \boldsymbol{Y}_{\mathrm{F}}) = H_{\boldsymbol{Y}}(\boldsymbol{Y}_{\mathrm{I}}) - H_{\boldsymbol{Y}}(\boldsymbol{Y}_{\mathrm{F}}) > \tau$$

Human Evaluation: Formality

Formality is marked more strongly in Online Source Inference outputs than in MultiTask outputs



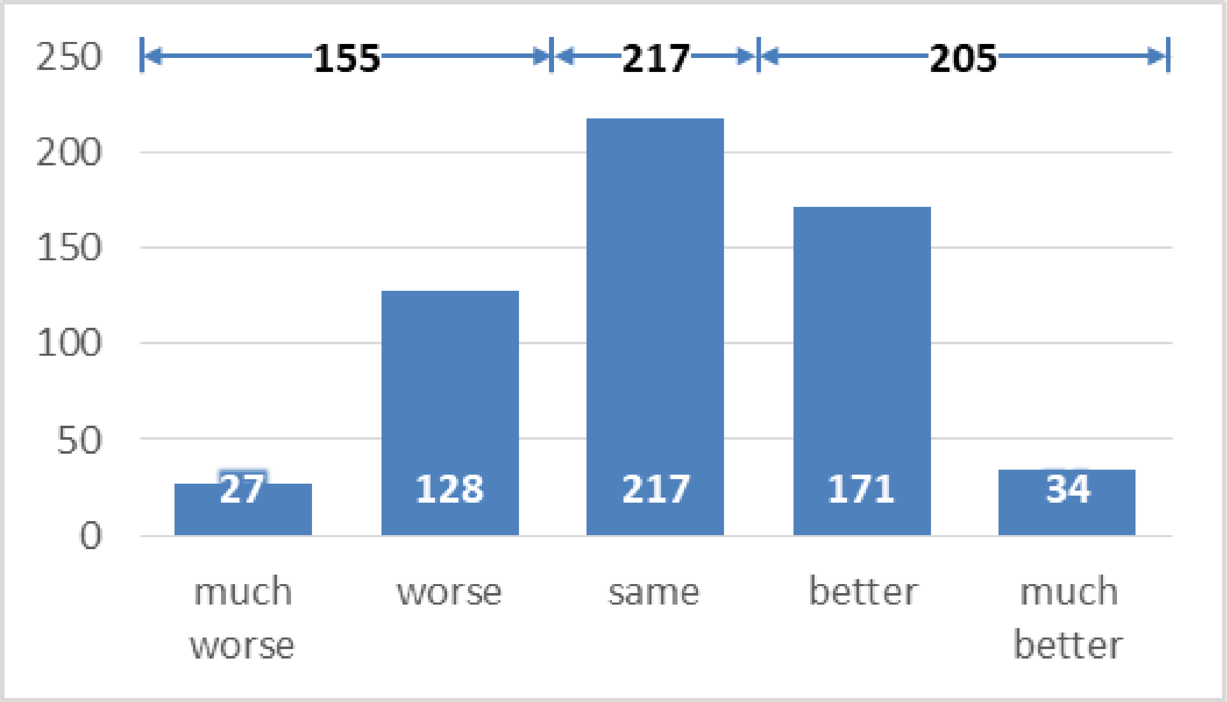
Informal translations



Formal translations

Human Evaluation: Meaning Preservation

Online Style Inference preserves the meaning of references better than Multitask



Style Divergences

Our new multitask formality transfer + MT model

Improves English formality transfer

Can produce distinct formal/informal translations of same input

Introduces more formality rewrites while preserving meaning, esp. with synthetic supervision

Style Divergences

Formality Style Transfer Within and Across Languages with Limited Supervision. Xing Niu, PhD Thesis 2019.

Multi-task Neural Models for Translating Between Styles Within and Across Languages. Xing Niu, Sudha Rao & Marine Carpuat. COLING 2018.

A Study of Style in Machine Translation: Controlling the Formality of Machine Translation Output. Xing Niu, Marianna Martindale & Marine Carpuat. EMNLP 2017.

github.com/xingniu/multitask-ft-fsmt

Semantic Divergences

Reference Divergences

Style Divergences

From Parallel Text to Machine Translation

Modeling divergences between reference & predictions improves NMT

$$\{(f_1, e_1), (f_2, e_2), \dots (f_N, e_N)\}$$

Detecting semantic divergence helps NMT training

$$e^* = \mathrm{argmax}_e \, p(e|f;\theta)$$

NMT can tailor output style while preserving input meaning

From Parallel Text to Machine Translation

How can we design training to best exploit available data?

$$\{(f_1, e_1), (f_2, e_2), \ldots (f_N, e_N)\}$$

What properties of training samples matter for training?

$$e^* = \mathrm{argmax}_e \, p(e|f;\theta)$$

Can we recast MT
as a language generation task?

Semantic, Stylistic & Other Data Divergences in Neural Machine Translation

Marine Carpuat

marine@cs.umd.edu

PhD student co-authors

Marianna Martindale

Xing Niu

Yogarshi Vyas

Aquia Richburg

Weijia Xu

Qualitative Analysis

Type	Informal translation	Formal translation
Filler	**And** I think his wife has family there.	I think his wife has family there.
Completeness ▼		
Quotation	The gas tax is simply not sustainable, said Lee.	"The gas tax is simply not sustainable," said Lee.
Yes-No	You like shopping?	**Do** you like shopping?
Subject	Sorry it's my fault.	**I'm** sorry it's my fault.
Article	Cookies where I work.	**The** cookies where I work.
Relativizer	Other stores you can't buy.	The other stores **where** you can't buy.
Paraphrasing ▼		
Contraction	I think **he'd** like that, but **we'll** see.	I think **he would** like that, but **we will** see.
Possessive	**Fay's innovation** perpetuated over the years.	**The innovation of Fay** has perpetuated over the years.
Adverb	I **told you already**.	I **already told you**.
Idiom	Hi, **how's it going**?	Hi, **how are you**?
Slang	You **gotta** let him digest.	You **have to** let him digest.
Word-1	Actually my **dad**'s some kind of technician so he understands, but my **mom**'s very old.	In fact, my **father** is some kind of technician so he understands, but my **mother** is very old.
Word-2	**Maybe** a little more in **some** areas.	**Perhaps** a little more in **certain** areas.
Word-3	It's **really necessary** for our nation.	This is **essential** for our nation.
Phrase-1	Yeah, **me neither**.	Yeah, **neither do I**.
Phrase-2	I think he's moving to California **now**.	I think he is moving to California **at the moment**.
Phrase-3	It could be **a Midwest thing**.	This could be **one thing from the Midwest**.

Intrinsic Evaluation: ConvNet trained on synthetic examples performs best

Divergence Detection	OpenSubtitles							Common Crawl						
Approach	+P	+R	+F	-P	-R	-F	Overall F	+P	+R	+F	-P	-R	-F	Overall F
Sentence Embeddings	65	60	62	56	61	58	60	78	58	66	52	**74**	61	64
MT Scores (1 epoch)	67	53	59	54	68	60	60	54	65	59	17	11	14	42
Non-entailment	58	78	66	53	30	38	54	73	49	58	48	72	57	58
Non-parallel	70	83	76	61	42	50	66	70	83	76	61	42	49	67
Semantic Dissimilarity	**76**	**80**	**78**	**75**	**70**	**72**	**77**	**82**	**88**	**85**	**78**	69	**73**	**80**