Tuning Neural MT

Guido Zarrella

MITRE Corporation

Guido Zarrella **Ivan Young** **Becky Marvin**

Outline

- **Tuning MT: when the system you have isn't the system you need**
- **Neural MT tuning methods differ from those for Statistical MT**
- **Genre or Domain matters (a lot):**
  - In-genre test: BLEU = **25.6**
  - Out-of-genre test: BLEU = **7.5** (**-18.1**)
- **You care about NMT tuning because…**
  - Tuned w/ monolingual data only: BLEU = 10.3 (**+2.8**)
  - Trained on a small parallel set: BLEU = 13.5 (**+6.0**)
  - Tuned (transfer learning): BLEU = 15.0 (**+7.5**) to 16.9 (**+9.4**)

Tuning a system you have, to get the system you need

Tuning a system you have, to get the system you need

Tuning Machine Translation

In SMT, tuning involves learning a **weighted combination** of scoring features output by trained components: **translation tables, language models, reordering models**, …

For example: Minimum Error Rate Training (MERT)
or Margin-infused Relaxed Algorithm (MIRA)

Tuning Statistical Machine Translation

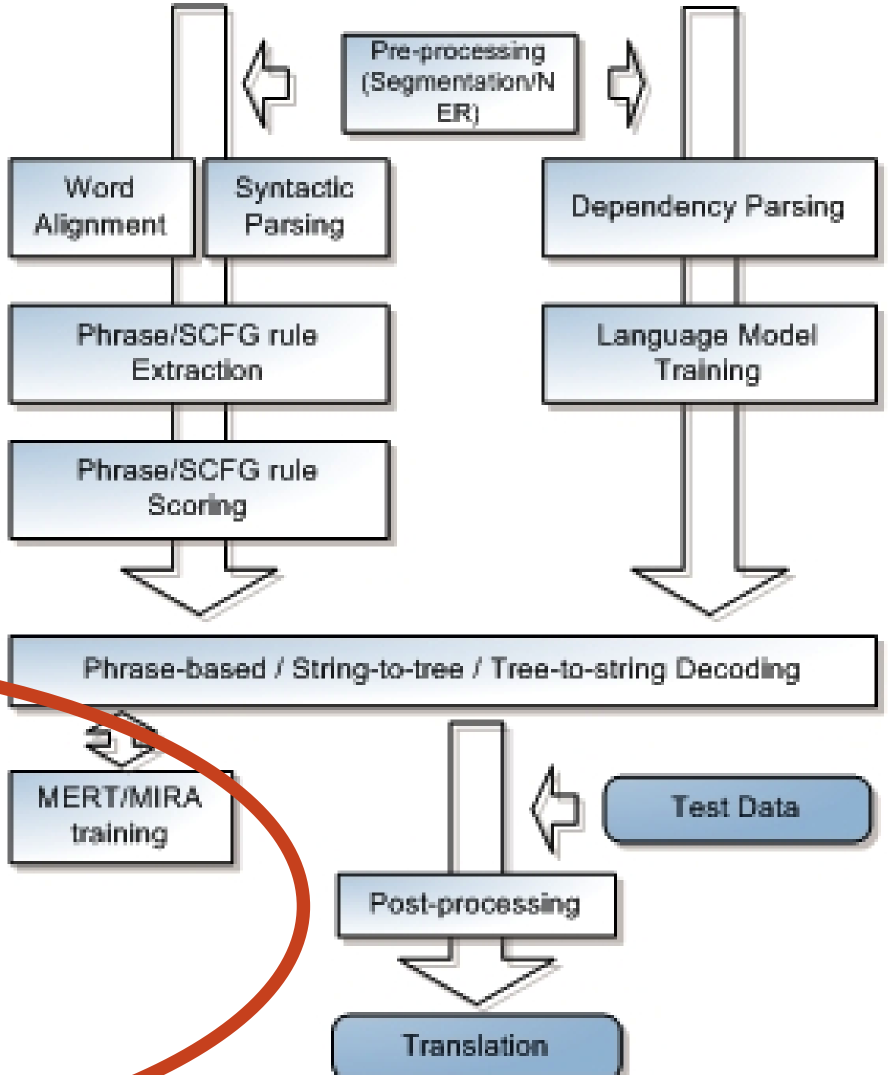


Image Credit: NLP Group at Northeastern University, China

Pair	System	untuned	MERT-tuned
fr-en	WMT-SMALL	28.0	29.2 (0.2)
	WMT-LARGE	29.4	32.5 (0.1)
de-en	WMT-SMALL	25.0	25.3 (0.1)
	WMT-LARGE	26.6	26.8 (0.2)

SampleRank Training for Phrase-Based Machine Translation
Barry Haddow, Abhishek Arun, Philipp Koehn 2011

Need for Domain Adaptation

Newswire source

目前日本有关方面已经派出三只 巡逻艇, 协同韩国方面在出事水域开展搜寻遇难者的工作.

Semiconductor source

利用在线应力测试技术表征了掺入Pt后对镍硅化物薄膜应力性质的影响.

Human translation

Currently, Japanese authorities have three dispatched patrol boats to coordinate with the South Koreans in searching for the victims in the area of the incident.

The effect of Pt doping on the stress in the nickel silicide film has been characterized using an in-situ stress measurement.

Machine translation

Japan has dispatched three patrol boats to the area, in coordination with the South Koreans to search for the victims in the area of the incident work

Stress tests use online technology characterized by incorporation of Pt on nickel silicide films nature of the stress

Quite poor on novel domains

Need for Domain Adaptation

System	Description	Score (BLEU) Semi-conductor	Chem-bio
L	Stand-alone product, statistical	9.4	9.7
S	Stand-alone product, rule-based	11.2	11.9
G	Web-based, statistical	15.1	22.8*
MITRE	Statistical	16.1	17.9

MITRE

Neural MT

"With the exception of fr-es and ru-en the neural system is always comparable or better than the phrase-based system."

Is Neural Machine Translation Ready for Deployment? A Case Study on 30 Translation Directions

Marcin Junczys-Dowmunt, Tomasz Dwojak, Hieu Hoang

Neural Machine Translation

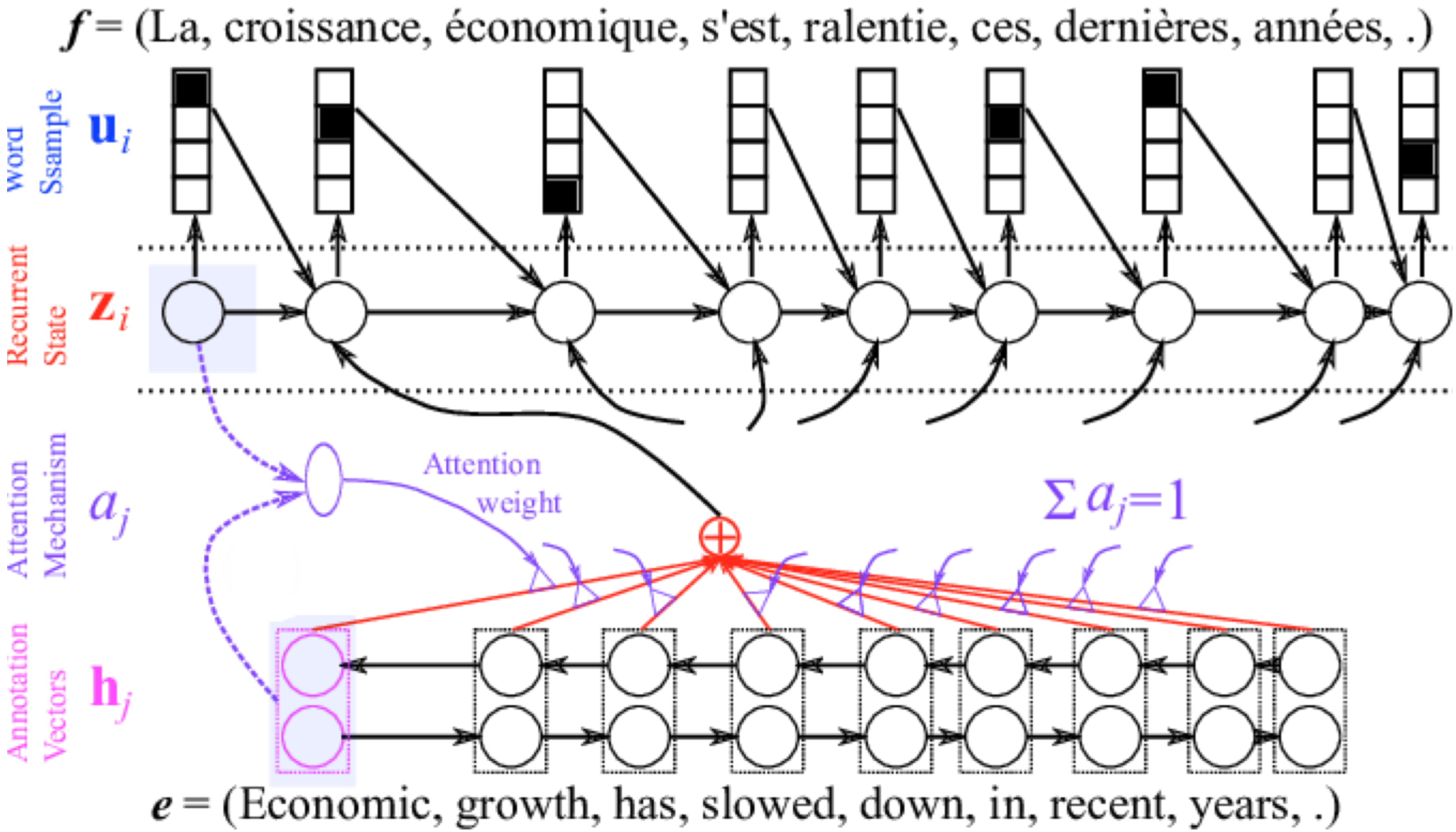


Image credit: Kyunghyun Cho: https://devblogs.nvidia.com/parallelforall/introduction-neural-machine-translation-gpus-part-3/

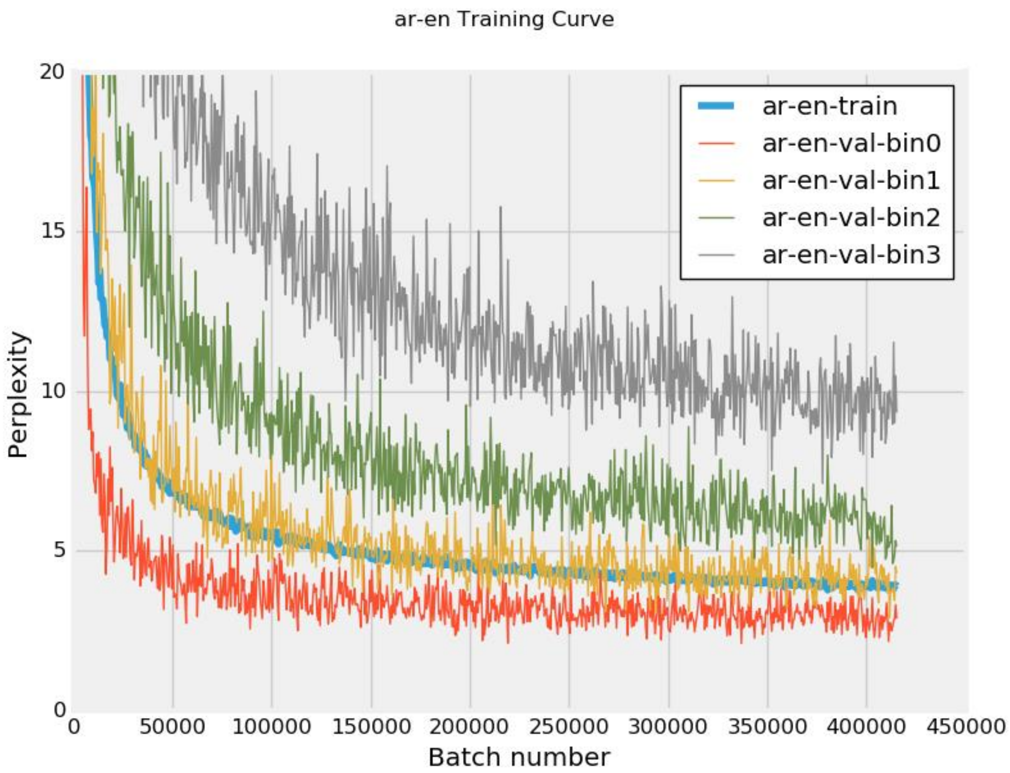
Subtitle Corpus for Discourse

Pierre Lison and Jörg Tiedemann, 2016, *OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles.* In Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016).

language	files	tokens	sentences	af	ar	bg	bn	br	bs	ca	cs	da	de	el	en	eo	es	et	eu	fa	fi	fr	g
af	32	0.2M	27.4k		6.2k	7.6k			1.8k		10.5k	6.0k	7.9k	11.6k	16.2k		12.6k	2.2k		2.1k	2.8k	7.3k	
ar	67,608	329.8M	60.8M	6.2k		16.2M	62.2k	13.0k	6.1M	0.3M	16.5M	7.5M	7.1M	15.3M	19.4M	19.4k	18.3M	6.9M	0.1M	3.0M	10.8M	14.4M	44
bg	90,376	523.4M	80.2M	7.7k	17.8M		60.7k	13.8k	7.5M	0.3M	21.1M	8.2M	8.9M	19.3M	26.4M	23.4k	24.8M	7.7M	0.1M	2.8M	13.2M	18.5M	48
bn	76	0.6M	0.1M		64.1k	62.7k			36.6k	3.1k	61.1k	58.2k	54.7k	58.5k	69.3k		65.8k	56.5k	3.1k	44.8k	56.1k	59.3k	
br	32	0.2M	23.1k		13.3k	14.1k			2.7k	5.3k	14.5k	10.0k	7.5k	14.4k	17.7k	1.1k	15.6k	15.0k	0.7k	4.4k	8.1k	15.4k	0
bs	30,511	179.5M	28.4M	1.8k	12.2M	8.5M	37.7k	2.7k		0.1M	7.5M	3.7M	3.6M	7.3M	9.5M	7.4k	9.0M	3.5M	76.3k	1.3M	5.2M	6.8M	27
ca	711	4.0M	0.5M		0.3M	0.3M	3.2k	5.5k	0.1M		0.3M	0.2M	0.2M	0.3M	0.4M		0.4M	0.2M		96.2k	0.2M	0.3M	11
cs	125,126	715.3M	112.8M	10.7k	18.1M	24.7M	63.3k	14.8k	8.5M	0.4M		8.5M	9.3M	19.8M	27.5M	31.7k	25.9M	7.9M	0.1M	2.9M	13.7M	19.1M	68
da	24,079	162.4M	23.6M	6.1k	8.0M	9.3M	60.9k	10.1k	4.0M	0.2M	9.6M		4.9M	8.1M	9.4M	11.3k	9.1M	5.0M	87.7k	2.1M	7.9M	7.6M	28
de	27,742	186.3M	26.9M	8.0k	7.6M	10.0M	57.2k	7.7k	4.0M	0.2M	10.6M	5.4M		9.1M	11.5M	24.9k	10.8M	4.3M	75.7k	1.8M	6.9M	9.2M	52
el	114,230	683.1M	101.6M	11.8k	16.8M	22.3M	60.5k	14.6k	8.1M	0.3M	23.0M	9.1M	10.2M		25.6M	24.5k	24.5M	7.5M	0.1M	2.8M	13.1M	19.6M	66
en	322,294	2.5G	336.6M	16.7k	21.9M	31.6M	75.0k	18.5k	11.1M	0.4M	33.8M	11.0M	13.4M	30.4M		49.0k	40.0M	8.6M	0.2M	3.3M	16.8M	28.0M	0.1
eo	89	0.5M	79.3k		19.9k	24.3k		1.1k	7.6k		32.8k	11.7k	25.6k	25.2k	51.1k		38.6k	17.6k		5.1k	18.9k	28.3k	0
es	191,987	1.3G	179.2M	12.9k	20.3M	29.2M	69.1k	16.0k	10.2M	0.4M	30.7M	10.4M	12.4M	28.6M	50.1M	40.2k		8.3M	0.2M	3.1M	15.7M	25.8M	0.2
et	23,515	140.7M	22.9M	2.2k	7.5M	8.9M	58.6k	15.4k	4.0M	0.2M	9.2M	5.7M	4.8M	8.6M	10.3M	18.2k	9.6M		93.3k	1.9M	6.5M	6.9M	29
eu	188	1.4M	0.2M		0.1M	0.1M	3.3k	0.7k	80.9k		0.1M	93.2k	80.1k	0.2M	0.2M		0.2M	0.1M		43.1k	0.1M	0.1M	10
fa	6,469	44.3M	7.4M	2.1k	3.1M	2.9M	46.3k	4.4k	1.4M	0.1M	3.1M	2.2M	1.9M	3.0M	3.6M	5.2k	3.3M	2.1M	44.7k		2.4M	2.5M	21
fi	44,594	208.5M	38.7M	2.8k	11.5M	14.8M	57.9k	8.3k	5.7M	0.2M	15.3M	9.0M	7.6M	14.6M	19.2M	19.5k	17.7M	7.4M	0.1M	2.5M		12.5M	40
fr	105,070	672.8M	90.9M	7.5k	15.5M	21.3M	61.4k	16.3k	7.4M	0.3M	21.8M	8.5M	10.3M	22.2M	33.5M	29.1k	30.1M	7.8M	0.1M	2.7M	13.9M		93
gl	370	1.9M	0.2M		45.8k	49.7k		0.5k	28.0k	11.9k	71.3k	29.4k	54.5k	68.8k	0.2M	0.3k	0.2M	29.9k	10.5k	22.0k	40.8k	96.1k	

Arabic to English

- **Trained on 21 million conversational segments from movie subtitles**
  - 256 million training steps (sentences)
  - 19 days on K40 GPU
- **NMT BLEU = 25.6**
  - SMT BLEU = 25.3
- **Serialized as 536 MB model**
  - Deployable to laptops

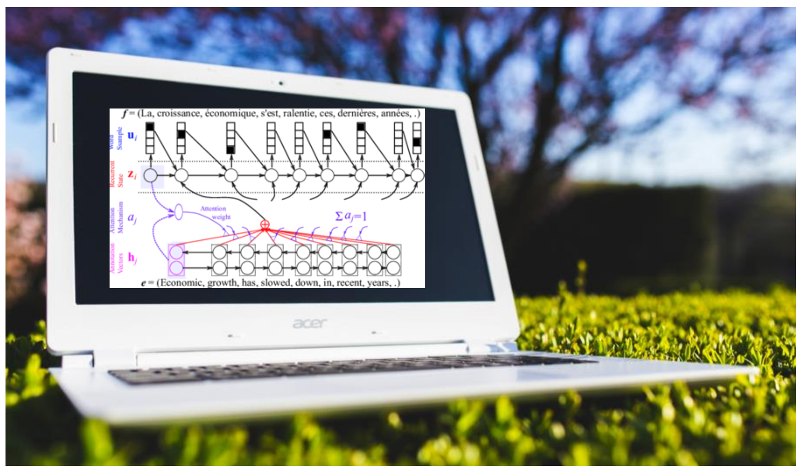


“26 BLEU”

OpenSubtitles Reference	NMT Output
people would think that he was the terrorist. right.	people say he was a terrorist . right .
- there's a boy in the cage.	- there ' s a boy in the cage .
we're just here to see our friend rigby, sir.	we ' re just here to see our friend , sir .
- glass is all over the floor. - somebody broke the stereo.	the glass is all around someone .
like nathan.	like a . .
let's get ice cream.	let ' s go get ice cream .
he's down checking a buoy in the channel.	he ' s out there asking for a consult .
oh, my god, please.	oh , god , please .
cervical lymph node has black flecks.	the black is a black .
you came for your uncle's wedding.	i came for your uncle ' s wedding .
yeah, and doctors say i should get more and more each day.	yeah , the doctors said i would remember more every day .

In a new domain

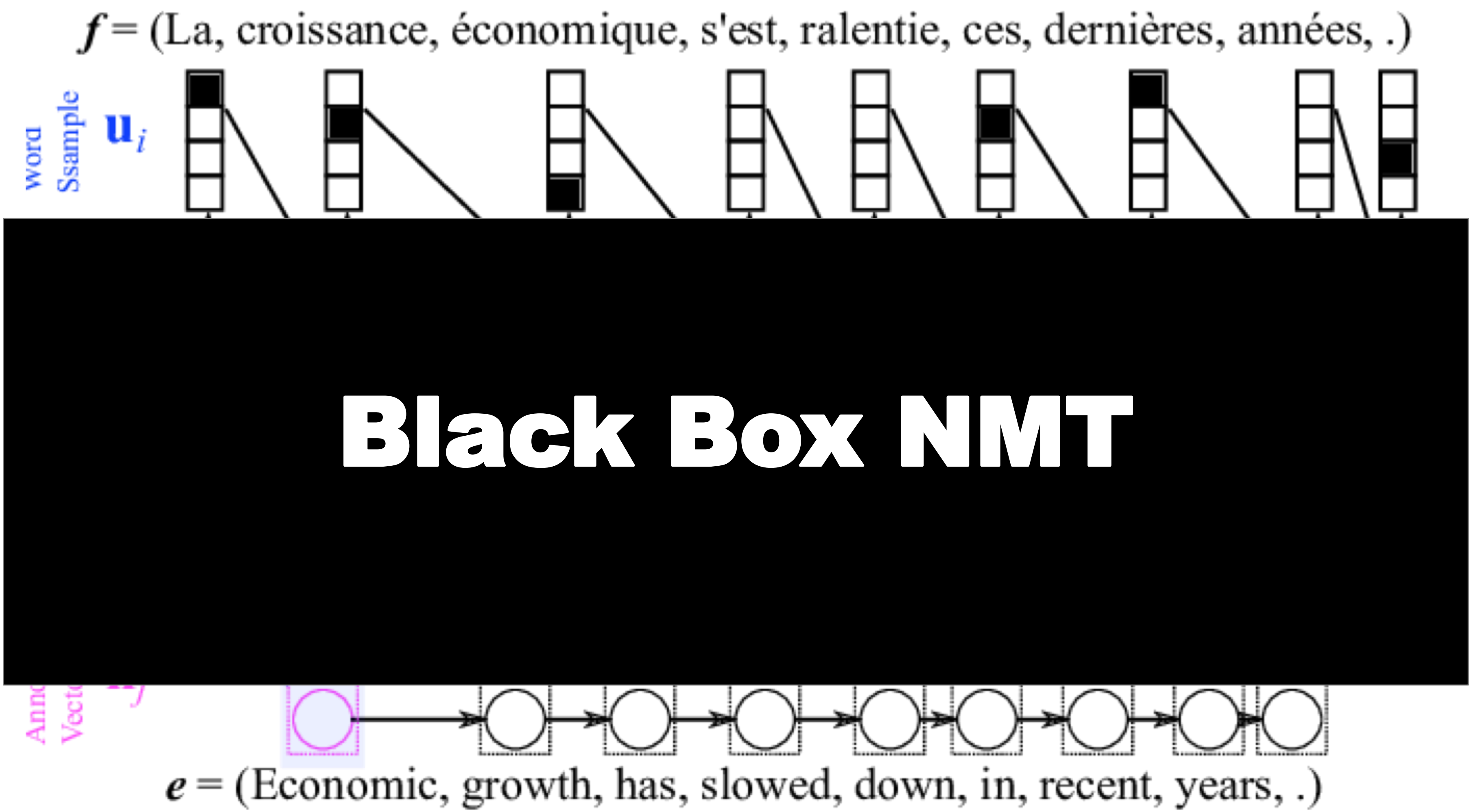
“tourism accounts for almost N % of the austrian gross domestic product .”



“the industry are nearly N , of the most common population .”

On Wikipedia:
BLEU = 7.4

Tuning NMT?

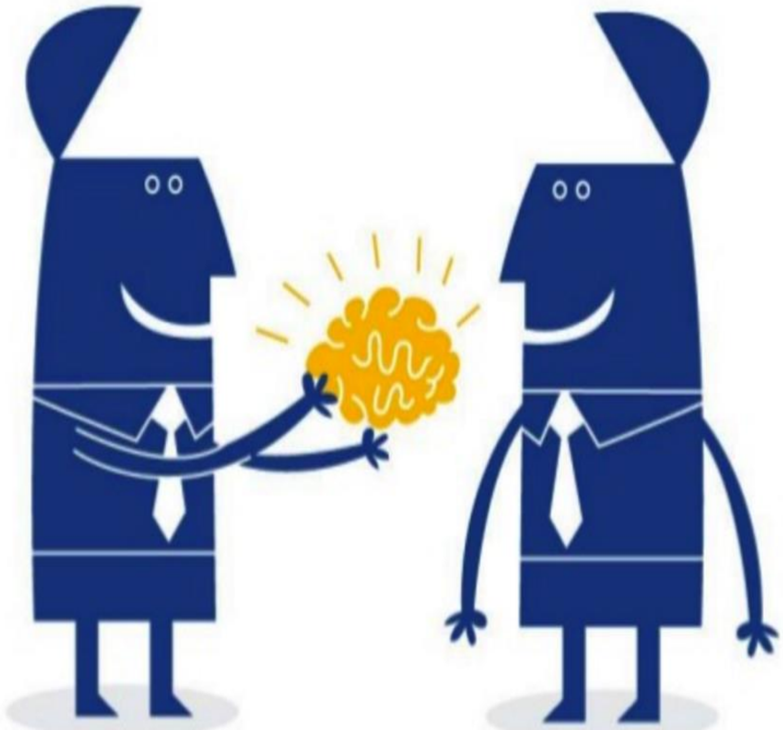


f = (La, croissance, économique, s'est, ralentie, ces, dernières, années, .)
Word
Ssample
u_i
Black Box NMT
e = (Economic, growth, has, slowed, down, in, recent, years, .)

Transfer Learning

- **Our core strategy is to employ transfer learning between deep neural networks pre-trained on massive datasets**

- **Knowledge gained in one context can be re-used to solve different but related problems**



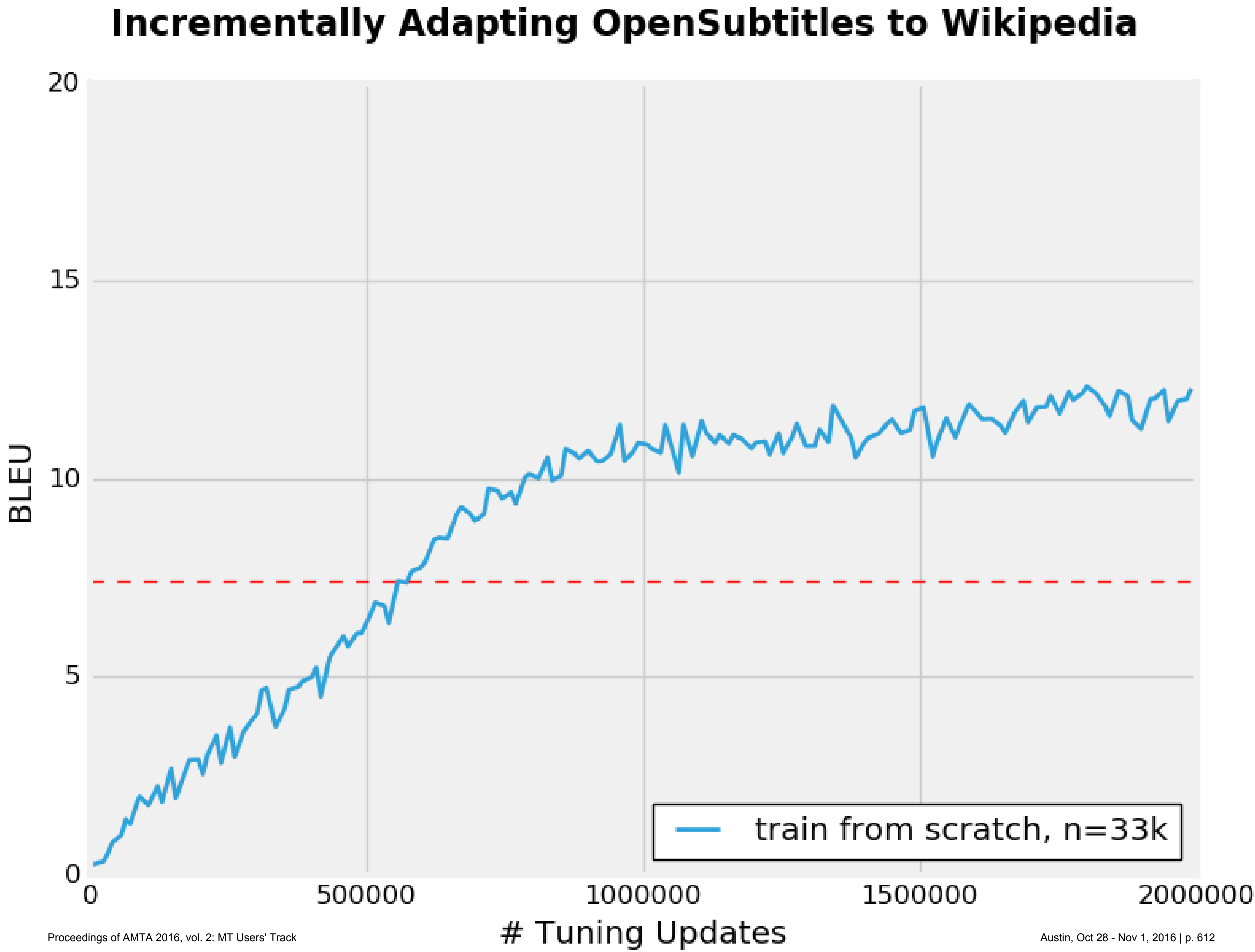
https://edpsychexperience.wordpress.com/2013/03/25/013112-learning-learning-transfer

Wikipedia Adaptation Experiments

- **Incremental training: we pick up where OpenSubtitles left off**
  – With tiny parallel tuning set (n=1024)
  – With small parallel training set (n=32768)
  – With full parallel training set (n=148136)
  – With varying amounts of in-domain monolingual data
  – With expanded vocabularies

- **About 22 minutes per 100k training updates**

Krzysztof Wołk and Krzysztof Marasek: Building Subject-aligned Comparable Corpora and Mining it for Truly Parallel Sentence Pairs., Procedia Technology, 18, Elsevier, p.126-132, 2014

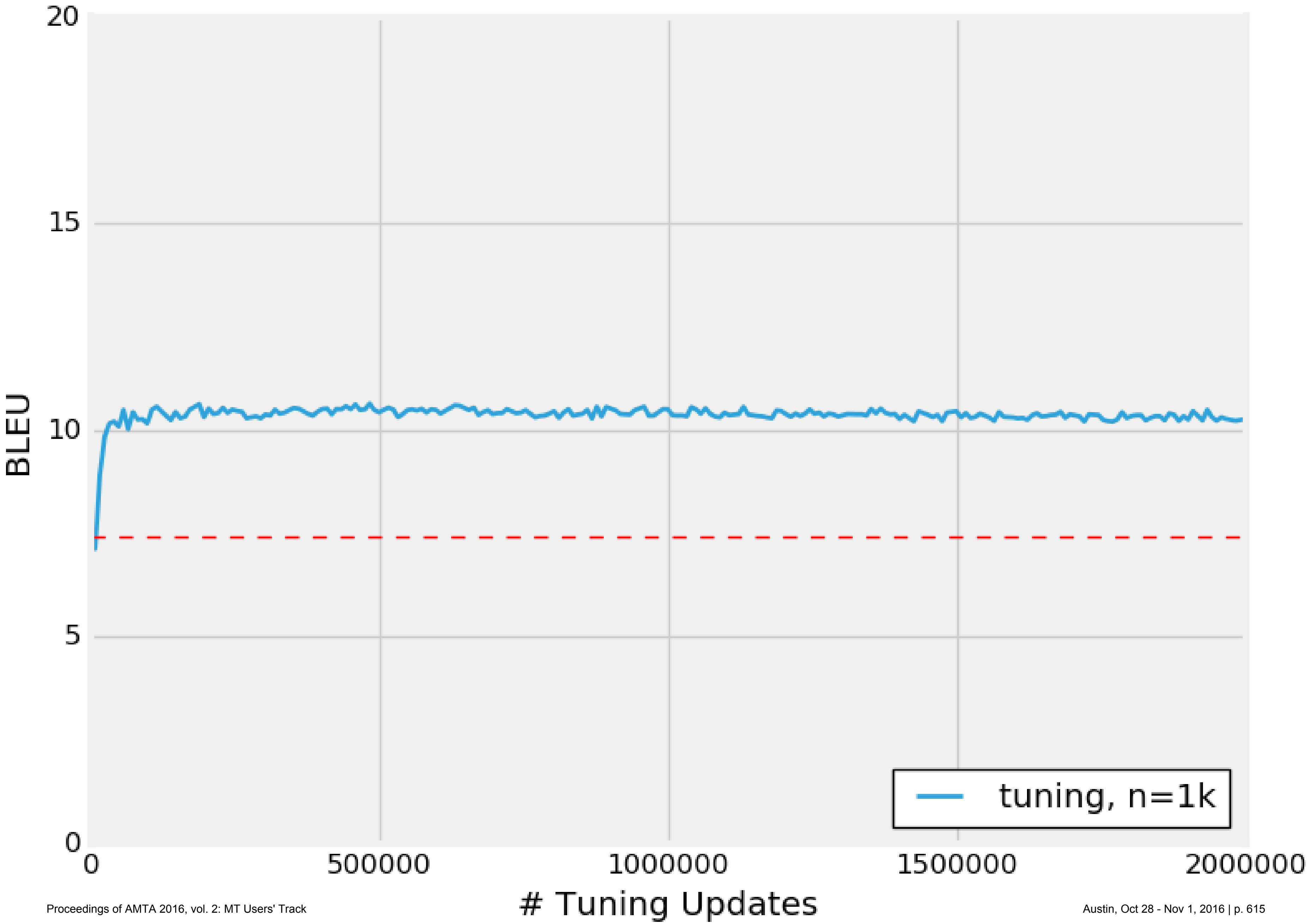


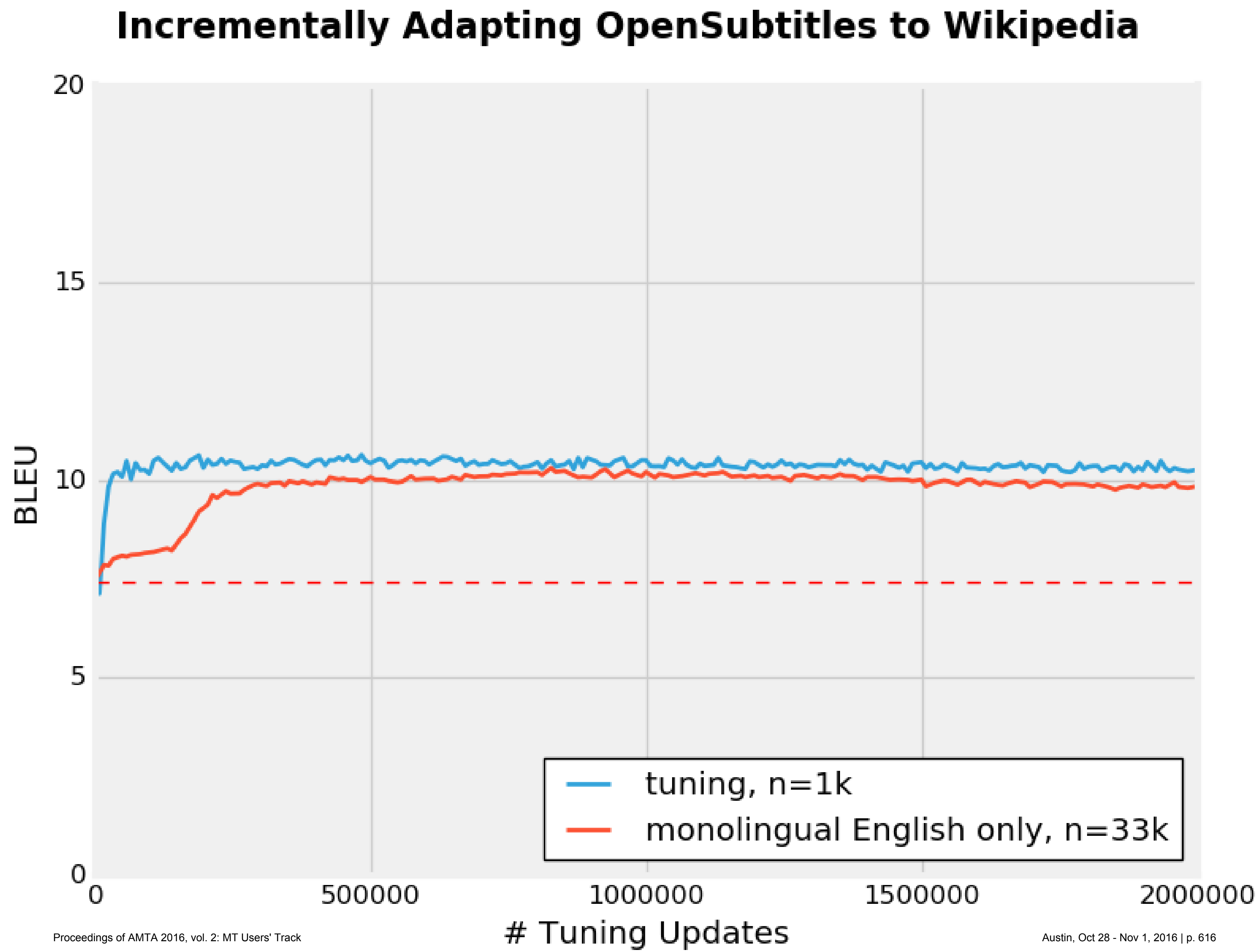
Incrementally Adapting OpenSubtitles to Wikipedia
20
15
10
5
0
BLEU
0
500000
1000000
1500000
2000000
Tuning Updates
train from scratch, n=33k

Incrementally Adapting OpenSubtitles to Wikipedia

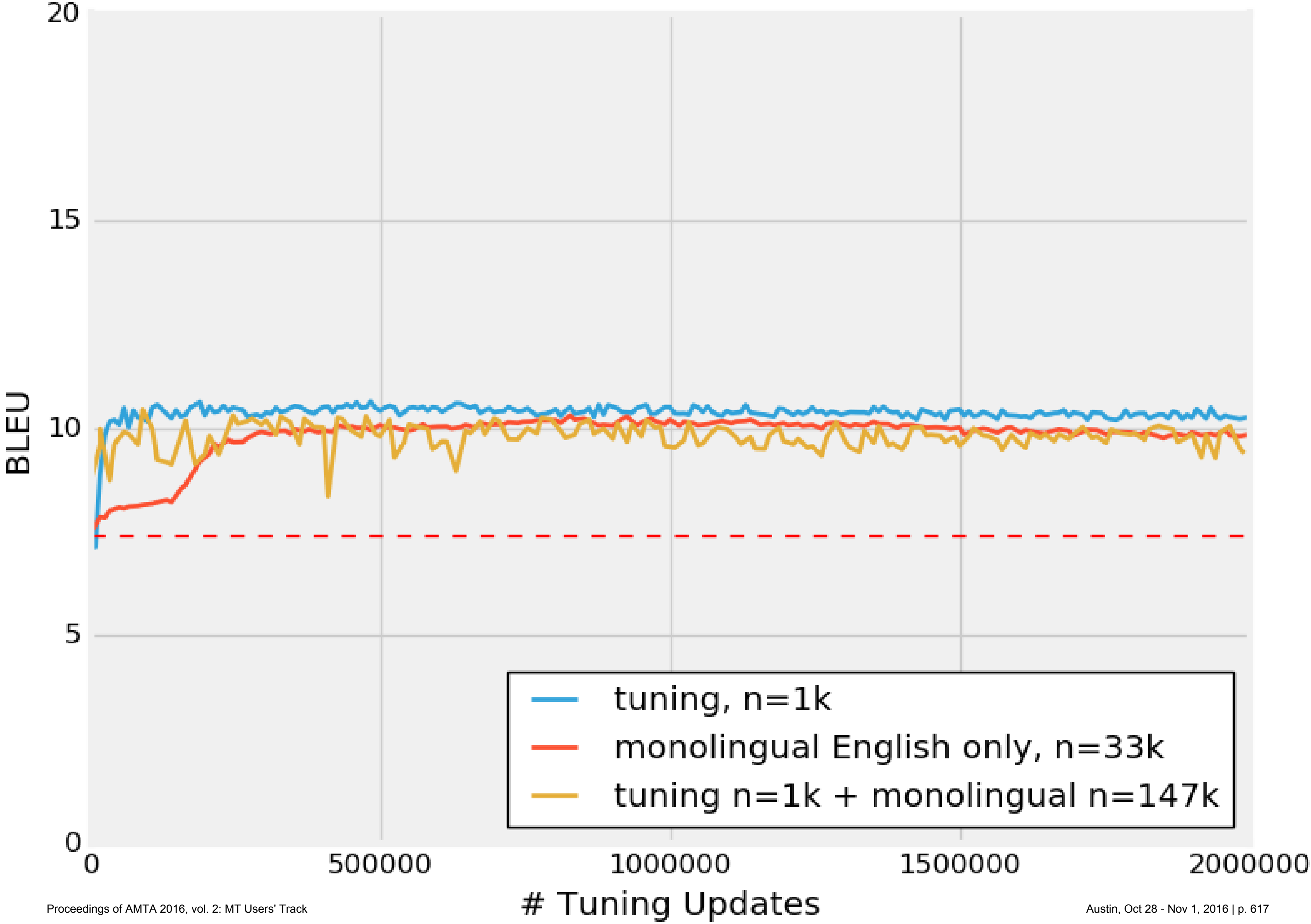
Incrementally Adapting OpenSubtitles to Wikipedia

Incrementally Adapting OpenSubtitles to Wikipedia




Incrementally Adapting OpenSubtitles to Wikipedia
20
15
10
5
0
BLEU
0
500000
1000000
1500000
2000000
Tuning Updates
tuning, n=1k
monolingual English only, n=33k

Incrementally Adapting OpenSubtitles to Wikipedia



Side by Side

Reference: tourism accounts for almost N % of the austrian gross domestic product .

Train from scratch, 33k: world is up for N % of the total reserves .

Untuned: the industry are nearly N , of the most common population .

1k tuning: tourism costs nearly N (of the most common population .

33k tuning: tourism often manifests approximately N % of the gdp .

… ensembling?

Results

- **Genre & domain matter (a lot)**
  - In-genre test: BLEU = **25.6**
  - Out-of-genre test: BLEU = **7.5** (**-18.1**)

- **Incremental training helps**
  - Trained, parallel in-domain: BLEU = **13.5 (+6.0)**
  - Tuned, parallel in-domain: BLEU = **15.0 (+7.5)** to **16.9 (+9.4)**

- **Monolingual data helps when parallel data is scarce**
  - Tuned, 33k monolingual in-domain: BLEU = **10.3 (+2.8)**
  - Tuned, 1k parallel in-domain: BLEU = **10.6 (+3.1)**

- **Expanding vocabulary doesn't increase BLEU (yet)**

MITRE

Conclusions

- **All parameters in a NMT system are tunable**
  - can create great diversity from one “well trained” seed system
  - … in minutes or hours, with little or no additional parallel data
- **Government use cases poised to benefit most**
  - Collect many partially trained systems on the shelf?
- **Still open question how to best create systems optimized for tuning**
- **Sharing models? Share training code too.**

Thank You

Guido Zarrella

jzarrella@mitre.org

@gzco