
SDL*
Turning NMT research into commercial products
Dragos Munteanu and Adrià de Gispert

SDL* helping big brands go global

- Founded in 1992
- 3800+ Employees
- 56 Offices
- 38 Countries
- 400 Partners
- 1500 Enterprise customers

78 of the top 100 global companies work with SDL

+10 BILLION words translated monthly

marketing campaigns	eCommerce	documentation	web, social media	analytics

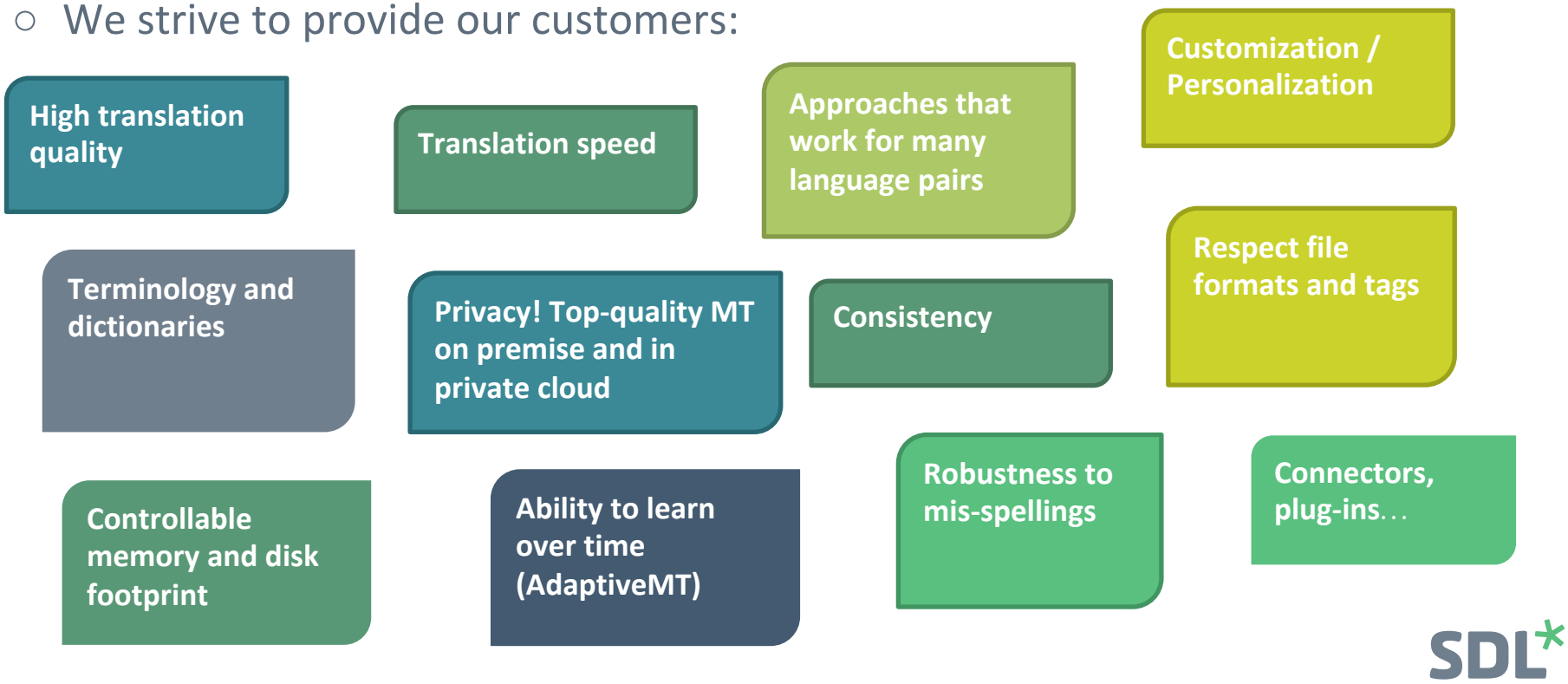
SDL Research – a long history in MT

- Research labs in Los Angeles (USA) and Cambridge (UK)
- Team members have published +100 on SMT and related tech
  - Bill Byrne, Abdessamad Echihabi, Dragos Munteanu, Gonzalo Iglesias, Eva Hasler, Adrià de Gispert, Steve DeNeefe, Jonathan Graehl, Wes Feely, Ling Tsou...
- Formerly Language Weaver
  - 15 years of leading expertise in SMT
  - major contributions (papers/patents) in phrase-based and string-to-tree MT, automata-based hierarchical MT, quality estimation, tuning, evaluation...
- Strong links with academia (University of Cambridge)
- Summer internships, industrial post-docs

SDL*

Our mission: Bring MT research results to products

- We strive to provide our customers:



SDL

SDL Secure Enterprise Translation Server

Data Security
- On premises/private cloud
- Used by gov't for 15 years

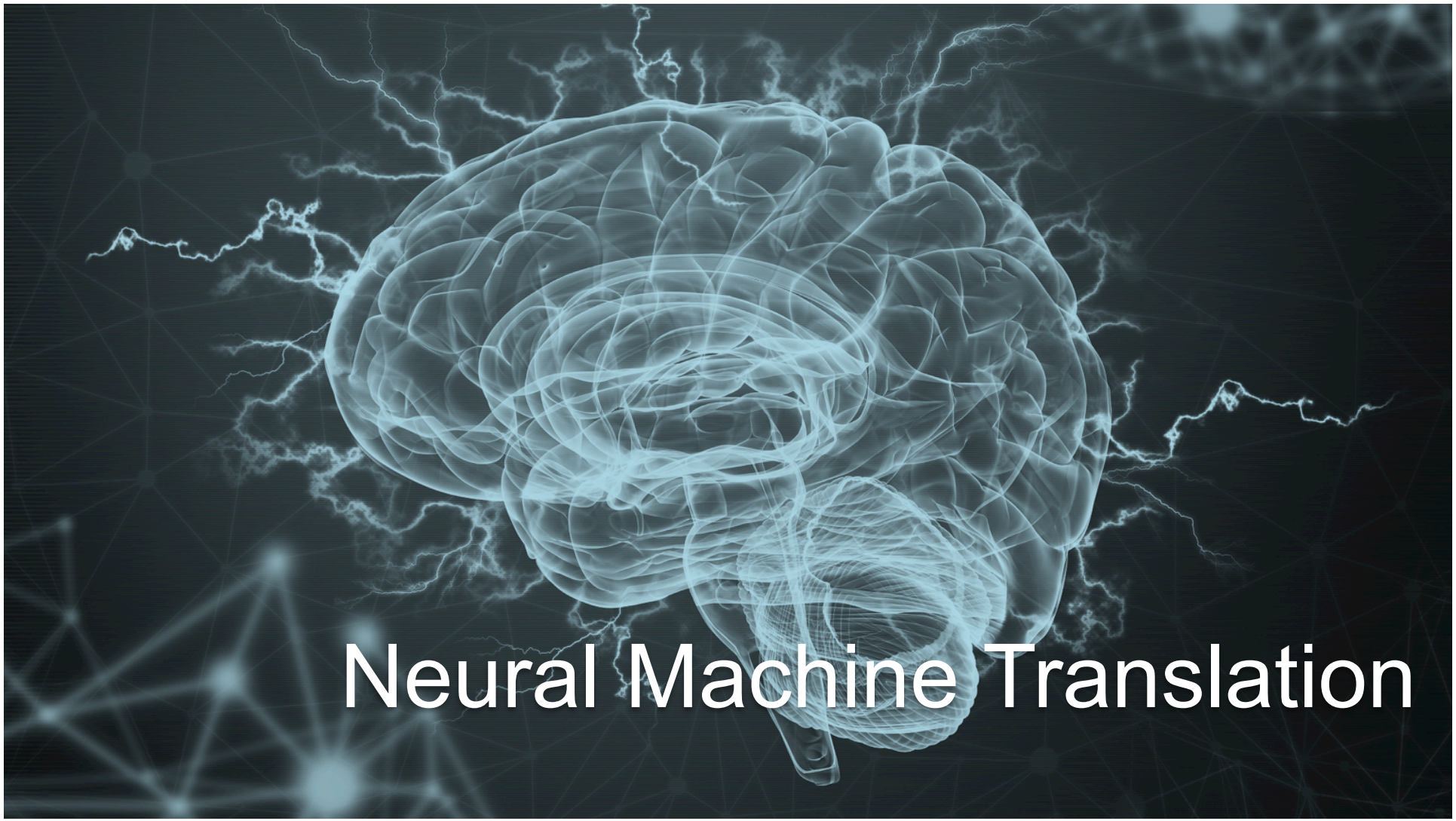
Quality / Customization
- Neural MT
- Custom MT out-of-the-box

Cost-effective scalability
- Elastic, optimized footprint
- Commodity hardware

Ease of Use / Integration
- deploys In hours
- MS plug-in & REST API

✓ 45 NMT engines currently available

SDL*


Neural Machine Translation

A paradigm shift

SMT

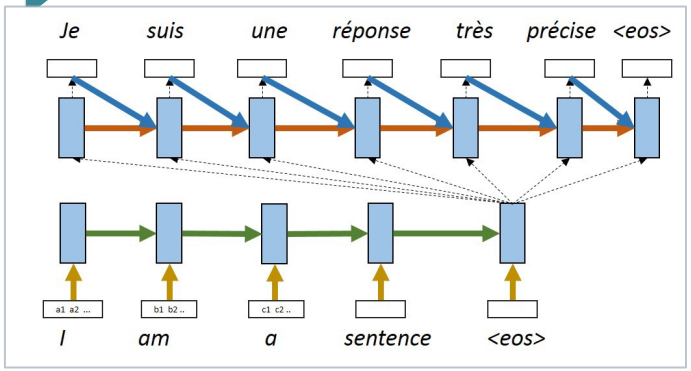
- Symbolic models
- Independence assumption (separate sub-problems)
- Maximum-likelihood estimation
- CPU-oriented training
- Source-side-guided decoding
- Large databases

Neural MT

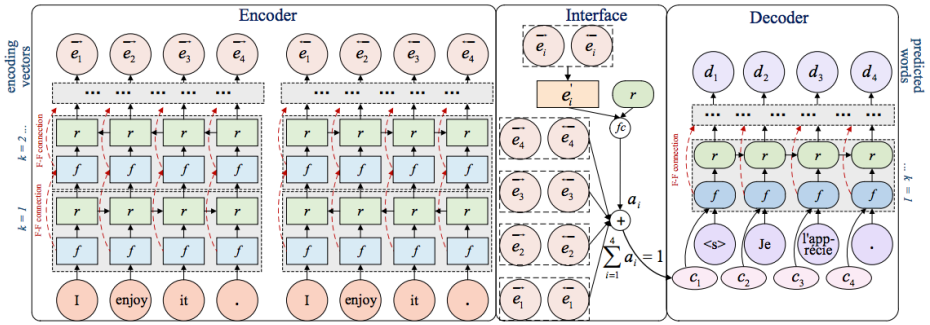
- Continuous-space models
- Single end-to-end model
- Discriminative training
- Reliance on GPUs
- Target-side-guided decoding
- Smaller compact models

SDL*

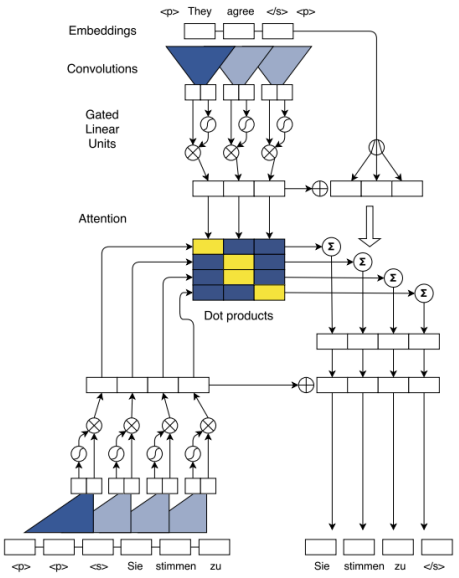
Better translation models



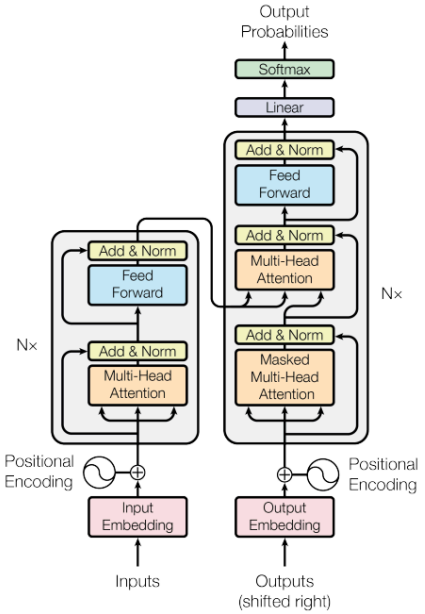
[Sutskever et al.'14] [Bahdanau et al.'15]



[Zhou et al.'16]

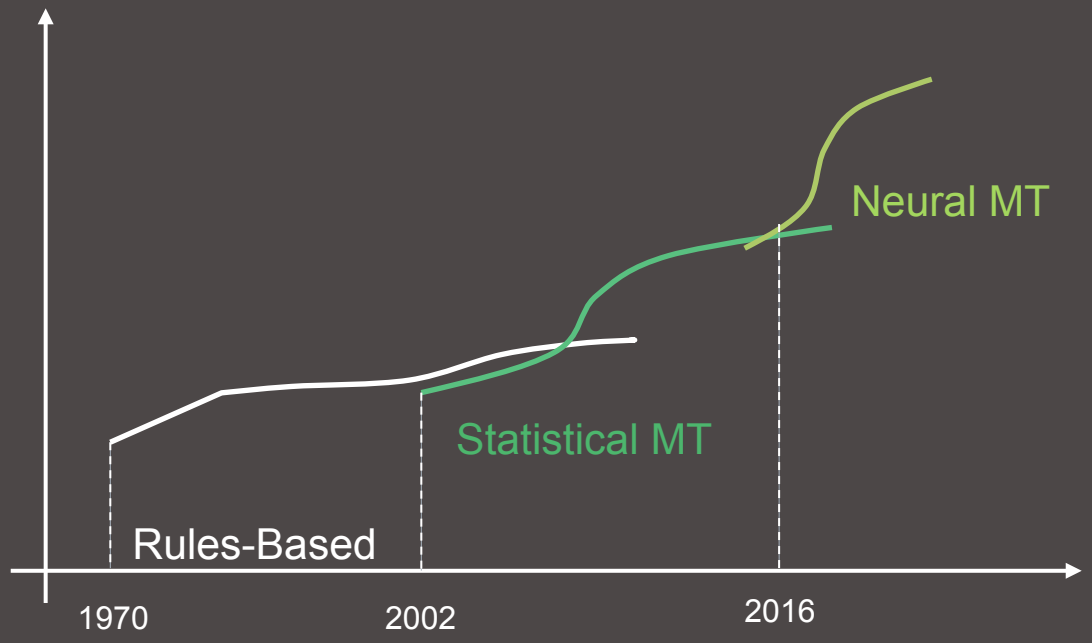


[Gehring et al.'17]

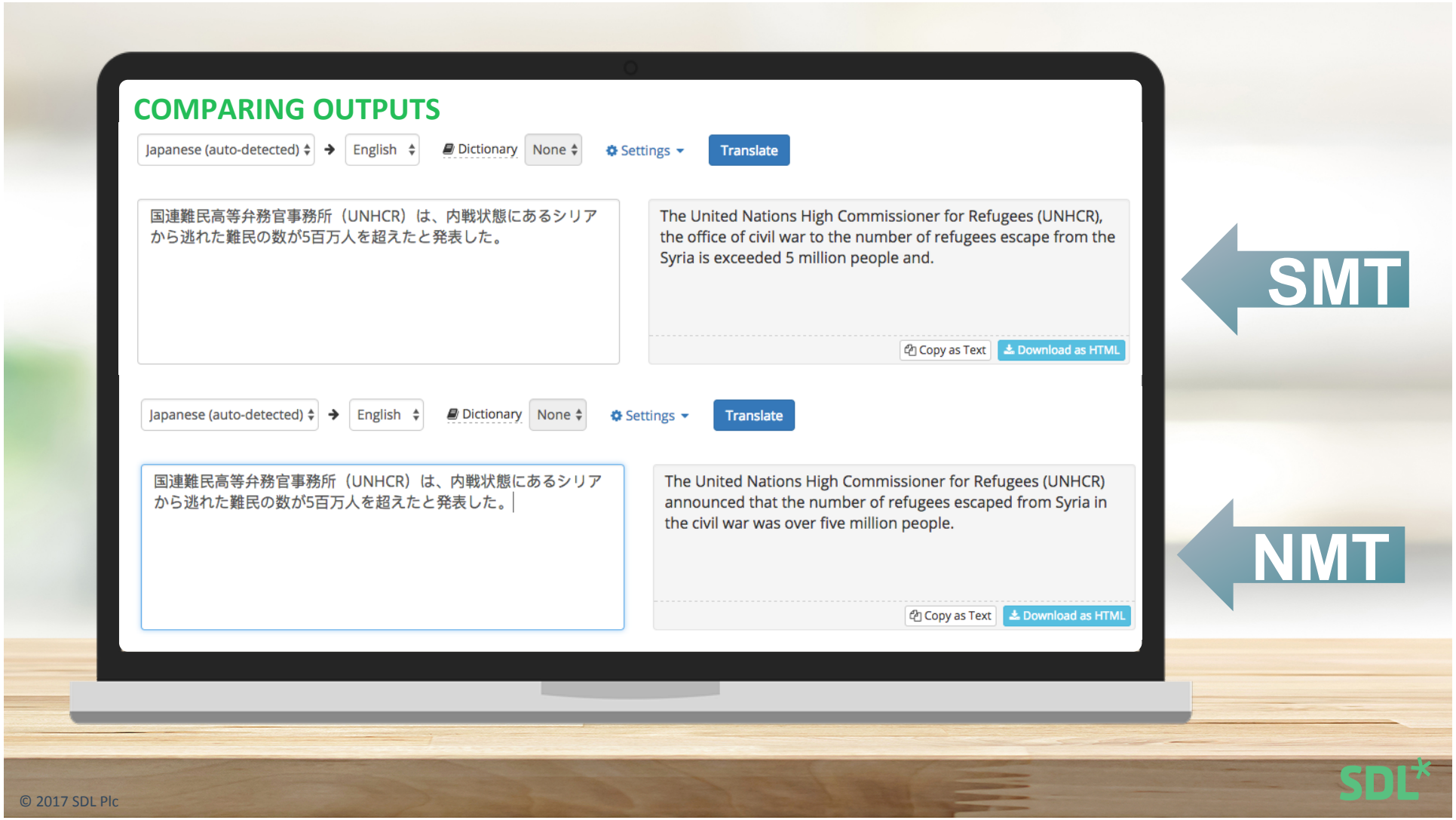


[Vaswani et al.'17]

Better BLEU scores



WAT	Jpn-Eng	Eng-Jpn
2014	23.8	35.0
2015	25.4	35.8
2016	27.6	36.2
2017	28.4	41.5
	+4.4 !!	**+6.5 !!**


COMPARING OUTPUTS
Japanese (auto-detected)
English
Dictionary
None
Settings
Translate
国連難民高等弁務官事務所（UNHCR）は、内戦状態にあるシリアから逃れた難民の数が5百万人を超えたと発表した。
The United Nations High Commissioner for Refugees (UNHCR), the office of civil war to the number of refugees escape from the Syria is exceeded 5 million people and.
Copy as Text
Download as HTML
SMT
Japanese (auto-detected)
English
Dictionary
None
Settings
Translate
国連難民高等弁務官事務所（UNHCR）は、内戦状態にあるシリアから逃れた難民の数が5百万人を超えたと発表した。
The United Nations High Commissioner for Refugees (UNHCR) announced that the number of refugees escaped from Syria in the civil war was over five million people.
Copy as Text
Download as HTML
NMT

Observable quality improvement

国連難民高等弁務官事務所（UNHCR）は、内戦状態にあるシリアから逃れた難民の数が5百万人を超えたと発表した。

Office of the United Nations High Commissioner for Refugees (UNHCR) is in a state of civil war when the number of refugees who have escaped from Syria have exceeded 5 million people.

The United Nations High Commissioner for Refugees (UNHCR) announced that the number of refugees escaped from Syria in the civil war was over five million people.

- ✓ 30% improvement over SMT across all our productized engines

SDL*

But… is it **ALL** that good?

There are situations in which NMT fails

- When it fails, it fails **spectacularly**
  - unrelated fluent text
  - repetitions, neurobabble…
- MT user/customer expectations
  - "MT is not supposed to do this" !!!?!
  - "Can it support the features I need" ???

SDL*

Over-generation and ‘neurobabble’

There was no clear correlation between the measured mass density and the measured mass density, and neither experiment A or B.

The company will pay approximately EUR 600 million in fines, and the U.S. Department of Justice (SEC) to pay for approximately EUR 600 million, and the U.S. Department of Justice and the Justice Department of Justice (SEC) to reduce the amount of internal control of the board of directors of the board of directors of the board of directors…

SDL*

Over-generation and ‘neurobabble’

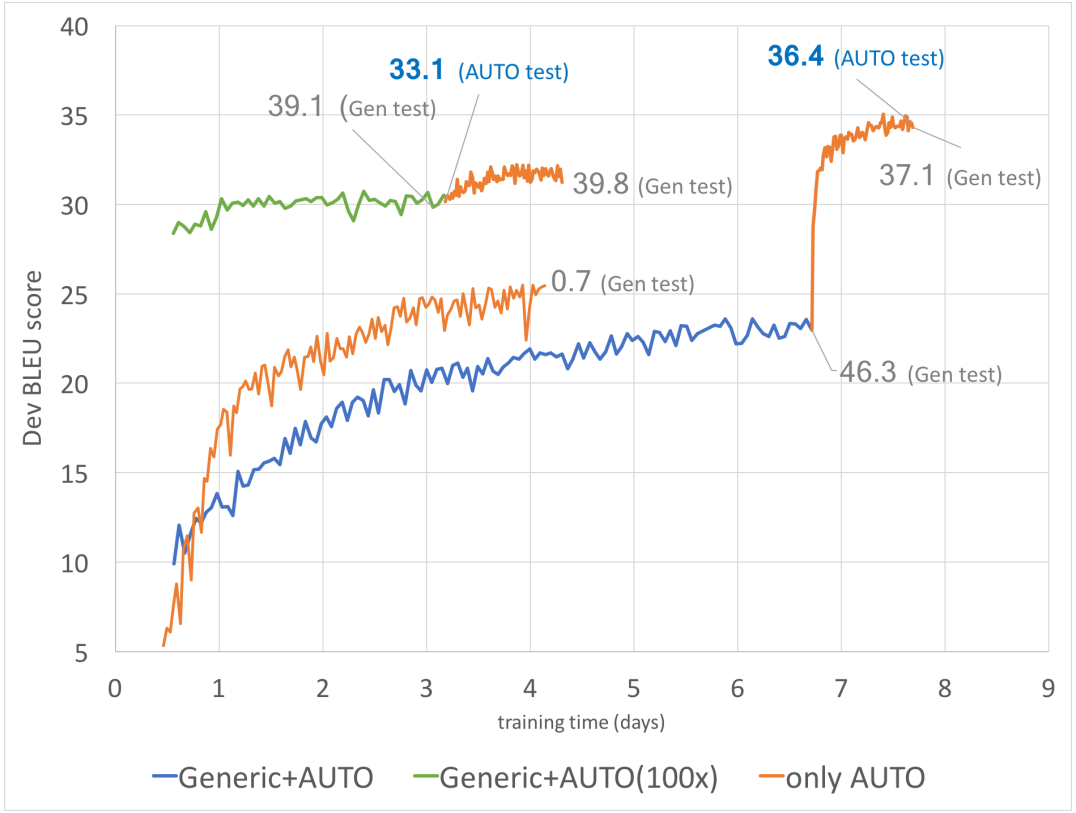
Data is EVEN MORE important

- New NMT models are better learners
  - A better fit to the training data
  - Relevant training data is key
  - Avoid babble and get huge gains!
- Domain adaptation/data selection

[Freitag and Al-Onaizan'16] [Chen et al.'17] [Britz et al.'17]
[Farajian et al.'17] [Van der Wees et al.'17] [Wang et al.'17]
…

SDL

Adapting neural models



Jpn-Eng corpus	# words
Generic	> 300M
Automotive	< 1M

Major improvements!

Challenge:

- Adapt to customer domain/data with minimal re-training
- Maintain high quality across domains

SDL

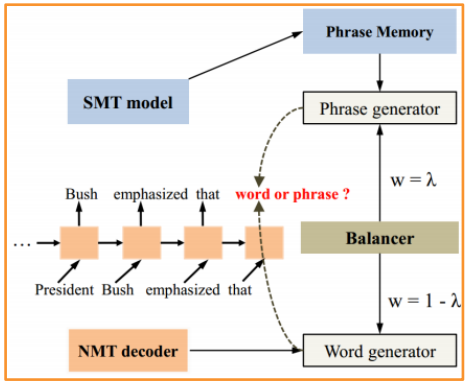
Lexical selection

- NMT models have freedom to produce any target word
  - Guided by source, not constrained
- SMT engines were good at lexical selection – can we leverage?
  - T-table, n-gram and phrase probabilities, memory-augmented models/search

[Arthur et al. EMNLP'16] [Stahlberg et al. EACL'17] [Wang et al.; Dahlmann et al.; Feng et al. EMNLP'17] [Zhang et al. IJCNLP'17] …

Input:	I come from Tunisia.
Reference:	チュニジア の 出身です。
	Chunisia no shusshindesu.
	(I'm from Tunisia.)
System:	ノルウェー の 出身です。
	Noruue- no shusshindesu.
	(I'm from Norway.)

[Arthur et al. EMNLP'16]



[Wang et al. EMNLP'17]

SDL*

NMT can use N-gram posterior probabilities

Best translation

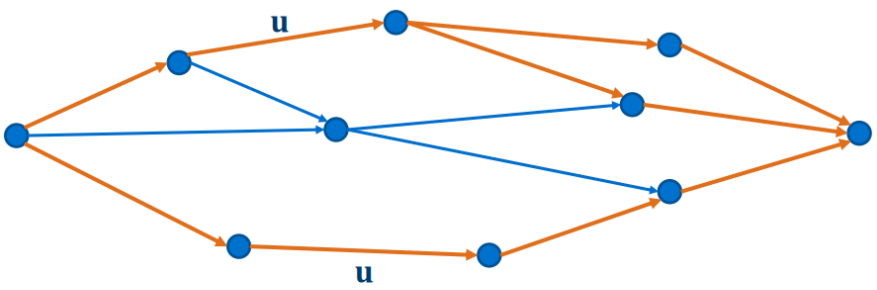
Number of n-gram $\mathbf{u}$ in translation $\mathbf{y}$.

Probability of n-gram $\mathbf{u}$ given the evidence space

$$\hat{\mathbf{y}} = \arg\max_{\mathbf{y} \in \mathcal{Y}_h} \Big(\underbrace{\Theta_0 |\mathbf{y}| + \sum_{\mathbf{u} \in \mathcal{N}} \Theta_{|\mathbf{u}|} \#_{\mathbf{u}}(\mathbf{y}) P(\mathbf{u}|\mathcal{Y}_e)}_{:= E_{SMT}(\mathbf{y})} \Big)$$

Hypothesis space of possible translations

Set of all n-grams



$P(\mathbf{u}|Y_e)$ = Sum of all orange path probabilities

$$\hat{\mathbf{y}} = \arg\max_{\mathbf{y}} \Big(E_{SMT}(\mathbf{y}) + \lambda \log P_{NMT}(\mathbf{y}|\mathbf{x}) \Big)$$

Evidence (~Risk) with respect to SMT lattice

Standard NMT translation score

BLEU scores

	Pure NMT	10k-best Rescoring	This Work (MBR-Based)
SMT Baseline[1]		22.2	
Single NMT (word)	22.5	24.5	**25.2**
6-Ensemble NMT (word)	25.0	25.4	**26.5**
3-Ensemble NMT (BPE)	25.9	25.1	**26.7**

Stahlberg et al. (EACL'17): "Neural Machine Translation by Minimising the Bayes-risk with Respect to Syntactic Translation Lattices"

SDL

But… are there guarantees?

- Control is **a must** for commercial success
- One very bad sentence can put off a customer
  - Back-off if needed

- Customers/Users expect certain ‘features’
  - Decoding speed, dictionary support, formatting constraints, Adaptive MT, …

SDL*

Dictionary support

"Zimra Games continues to innovate with the release next month of Coke Assault 3, which will satisfy the most demanding gamers."

English	German
Zimra Games	Zimra Games GmbH
Coke Assault 3	Coke Assault III
…	…

- Translation output **must** translate dictionary matches exactly – constrained search
- Easy for SMT decoders
- NMT beam decoder does not keep an alignment between source and target words

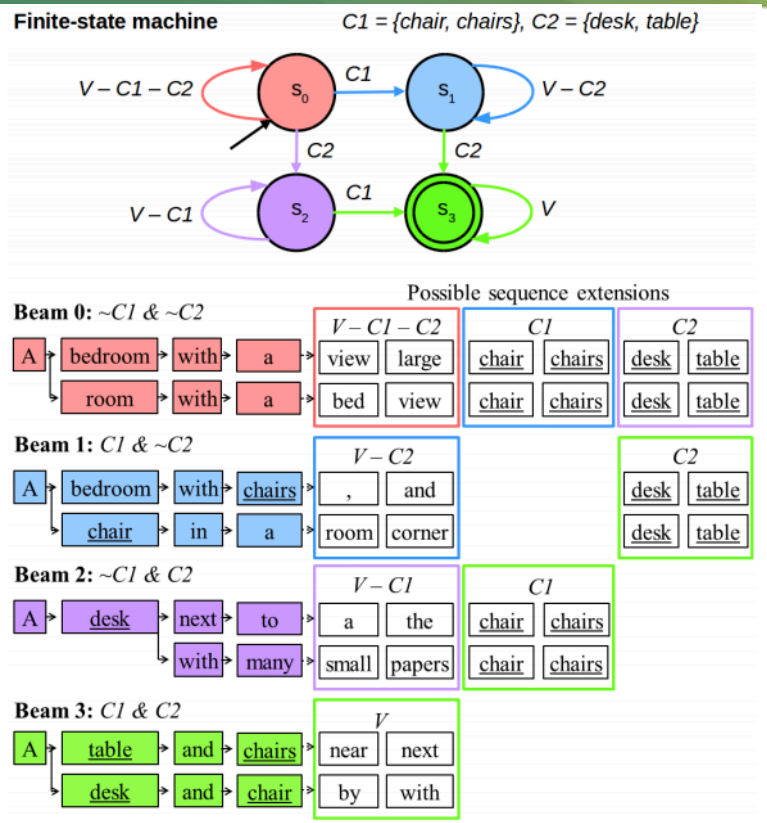
[Anderson et al. EMNLP'17]
[Hokamp & Liu ACL'17]
[Chatterjee et al. WMT'17]

SDL*

Dictionary support

Constrained search

- Build a finite-state acceptor with the target-side constraints
- Keep one separate stack per each acceptor state
- Output only hypotheses from the final acceptor state
- Constraints can be words or phrases



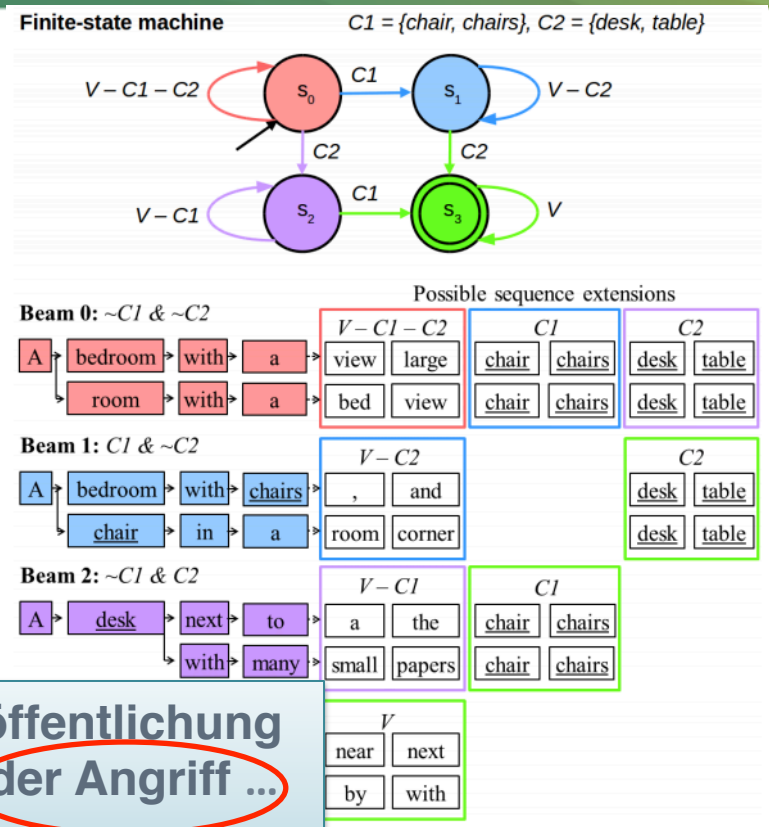
[Anderson et al. EMNLP'17]

SDL

Dictionary support

Challenges

- Computational complexity grows exponentially with the number of constraints
  - order is unknown
- Nothing prevents repeated decoding:



"**Zimra Games GmbH** setzt mit dem Veröffentlichung auf **Coke Assault III** im nächsten Monat der Angriff ...

SDL*

Entity constraints

"<B>Zimra Games</B> continues to innovate with the release <I>next month</I> of <B>Coke Assault <c=red>3</c></B>, which will satisfy the most demanding gamers."

- Decoder must also respect meta-tags
  - Key to support file formats used by MT users
- NMT model should not break sequential history
- Solutions require model specialization and/or decoding restrictions

SDL*

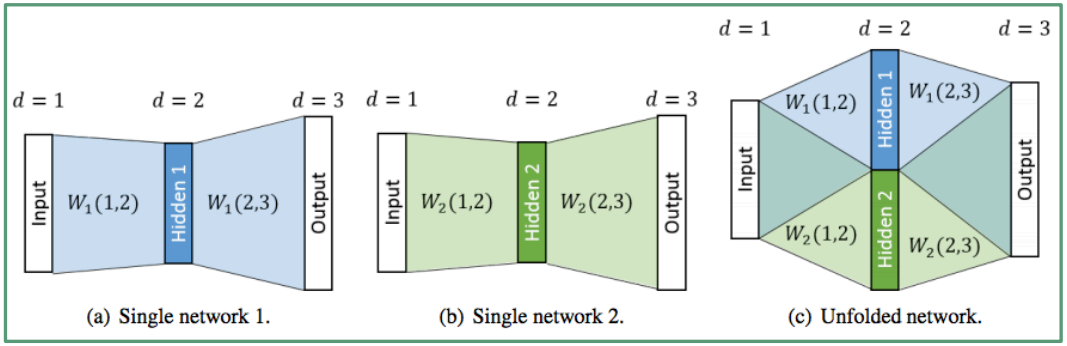
Decoding speed

- MT users are expected to certain translation speeds
  - Target speed varies, but well above research engines
- Goal is to provide best quality at desired speed
  - Speed vs quality trade-off
- NMT deployment scenarios
  - CPU only – hand-held devices, …
  - GPU
- NMT training speed also relevant

SDL*

Decoding speed vs quality trade-off (1)

- Model architecture
  - recurrent, convolutional, attentional…
  - number of parameters, layer precomputations…
  - Unfolding and shrinking ensembles



System	Words/Min. CPU	Words/Min. GPU	Size Factor	BLEU dev	BLEU test
Single	323.4	2993.6	1.00	20.8	23.5
2-Ensemble	163.7	1641.1	2 × 1.00	22.7	25.2
2-Unfold, shrunk embed.& attention	157.2	2592.2	1.77	22.7	25.1
2-Unfold, shrunk all except maxout	308.3	2961.4	1.05	22.4	25.3

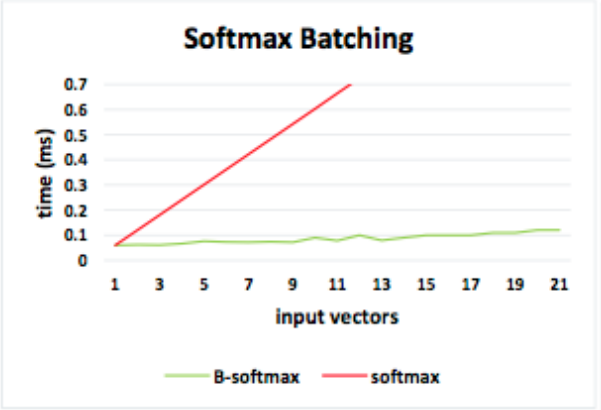
[Stahlberg and Byrne, EMNLP'17]

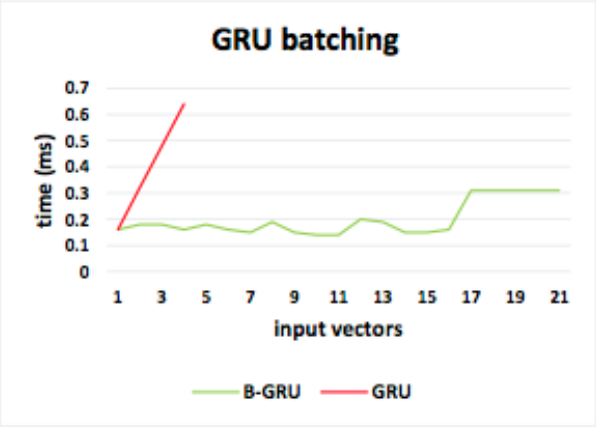
Stahlberg and Byrne (EMNLP'17): "Unfolding and Shrinking Neural Machine Translation Ensembles"

SDL*

Decoding speed vs quality trade-off (2)

- Hardware and Linear Algebra library
  - Type of GPU card
  - CPU-GPU communication
  - GPU usage
- Batching
  - standard in training





Decoding speed vs quality trade-off (3)

- Decoding parameters
  - beam size, early stopping…
- Reduced vocabulary softmax (CPU)
- Weight clipping in training
  - Low-precision inference

[Wu et al.'16] [Devlin, EMNLP'17] …

SDL

Thank you for your attention!

SDL

Software and Services for Human Understanding