Machine Translation

Info 159 — Fall 2017 — John DeNero

Translation Examples

Translation Task

- Language as input & language as output
- Input & output have roughly the same information content
- Conditional distribution over sentences is low entropy
- Lots of naturally occurring examples, but not much metadata
- Different objective functions for different applications

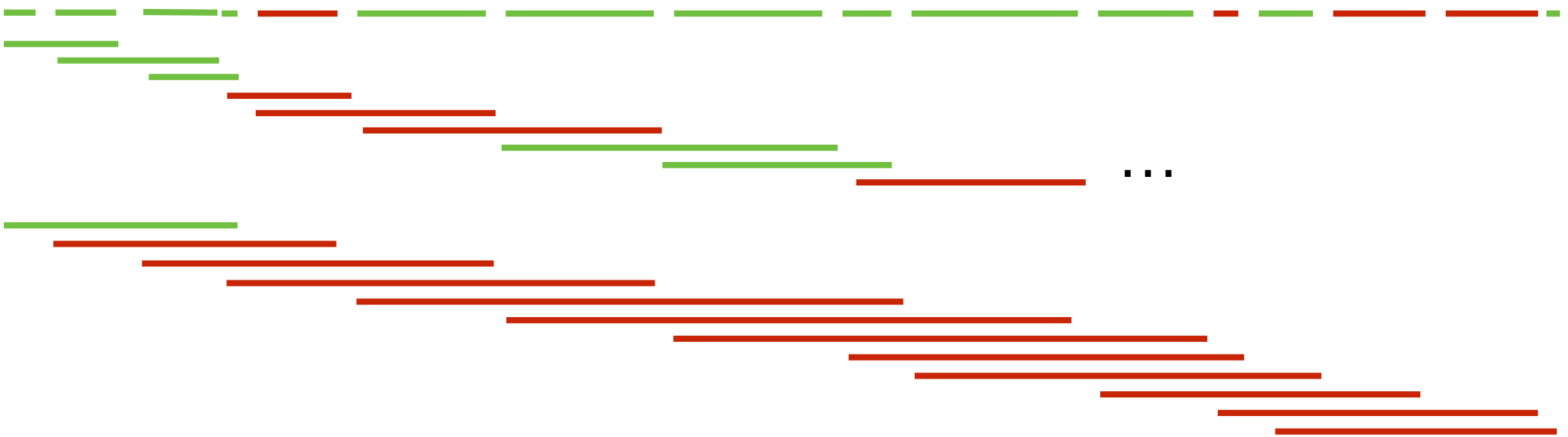
English-German News Test 2013 (a standard dev set)

```
Republican leaders justified their policy by the need to combat electoral fraud.

Die  Führungskräfte  der     Republikaner  rechtfertigen  ihre  Politik   mit   der     Notwendigkeit  ,  den  Wahlbetrug      zu  bekämpfen  .
 |          |         |          |             |         |      |      |    |    |          |        |  |    |      |         |      |     |
The  Executives      of the  republican    justify       your  politics  With  of the  need          ,  the  election fraud  to  fight     .

Republican leaders justify their policies with the need to fight electoral fraud.


However, the Brennan Centre considers this a myth, stating that electoral fraud is rarer in the United States than the number of people killed by lightning.

Allerdings  hält   das  Brennan  Center  letzteres   für  einen  Mythos  ,  indem  es  bekräftigt  ,  dass  der     Wahlbetrug      in  den  USA  seltener     ist  als  die  Anzahl  der     vom   Blitzschlag
    |       |      |     |       |       |          |    |     |      |    |    |     |         |    |     |    |      |         |    |    |    |     |       |       |       |       |
Indeed      keeps  the  Brennan  center  the latter  For  one    myth    ,  while  it  reaffirms   ,  that  of the  election fraud  in  the  USA  less common  is   as   the  number  of the  from  lightning stri

However, the Brennan Center thinks this is a myth by reiterating that electoral fraud in the US is less common than the number of people killed by lightning.


Indeed, Republican lawyers identified only 300 cases of electoral fraud in the United States in a decade.

Die  Rechtsanwälte  der     Republikaner  haben    in  10  Jahren  in  den  USA  übrigens    nur   300  Fälle  von   Wahlbetrug      verzeichnet  .
 |         |         |          |          |     |   |     |     |    |    |       |        |    |    |     |      |         |           |
The  Lawyers        of the  republican    to have  in  10  years   in  the  USA  by the way  just  300  cases  from  election fraud  listed      .

Incidentally, Republican lawyers have recorded only 300 cases of electoral fraud in the United States in 10 years.


One thing is certain: these new provisions will have a negative impact on voter turn-out.

Eins  ist  sicher    :  diese  neuen  Bestimmungen  werden  sich        negativ   auf  die  Wahlbeteiligung  auswirken  .
 |    |    |      |  |     |      |          |         |     |          |       |    |  |          |           |     |
one   is   for sure  :  these  new    provisions    become  themselves  negative  on   the  voter turnout    affect     .

One thing is certain: these new provisions will have a negative impact on turnout.


In this sense, the measures will partially undermine the American democratic system.

In  diesem  Sinne   untergraben  diese  Maßnahmen   teilweise  das  demokratische  System  der     USA  .
|     |       |         |        |          |          |       |         |         |       |   |    |   |
In  this    senses  undermine    these  activities  partially  the  democratic    system  of the  USA  .

In this sense, these measures partially undermine the democratic system of the United States.


Unlike in Canada, the American States are responsible for the organisation of federal elections in the United States.

Im      Gegensatz  zu  Kanada  sind  die  US-Bundesstaaten  für  die  Durchführung  der     Wahlen     in  den  einzelnen  Staaten  verantwortlich  .
|           |      |     |      |    |          |           |    |       |         |       |    |    |       |         |           |
in the  contrast   to  Canada  are   the  US States        For  the  execution    of the  elections  in  the  each      States   responsible     .

Unlike Canada, the US states are responsible for conducting elections in each state.


It is in this spirit that a majority of American governments have passed new laws since 2009 making the registration or voting process more difficult.

In  diesem  Sinne   hat  die  Mehrheit  der     amerikanischen  Regierungen  seit   2009  neue  Gesetze  verkündet    ,  die  das  Verfahren  für  die  Registrierung  oder  den  Urnengang  erschweren  .
|     |       |      |    |       |      |          |           |           |       |     |     |       |          |  |    |    |          |       |    |    |          |          |
In  this    senses  Has  the  majority  of the  American       governments  since  2009  new   laws    promulgated  ,  the  the  process    For  the  Registration   or    the  polls     aggravate   .

In this sense, the majority of American governments since 2009 have promulgated new laws that complicate the registration procedure or the polling process.


This phenomenon gained momentum following the November 2010 elections, which saw 675 new Republican representatives added in 26 States.

Dieses  Phänomen    hat  nach   den  Wahlen     vom   November  2010  an  Bedeutung   gewonnen  ,  bei  denen     675  neue  republikanische  Vertreter      in  26  Staaten  verzeichnet  werden  konnten  .
  |        |        |    |      |    |          |     |         |     |   |          |         |    |  |    |       |          |          |  |    |       |          |       |       |
This    phenomenon  Has  after  the  elections  from  November  2010  at  importance  won       ,  at   to those  675  new   Republican      representative  in  26  States   listed      become  could    .

This phenomenon has grown in importance following the November 2010 elections, which saw 675 new republican representatives in 26 states.
```

Evaluation with BLEU

In this sense, the measures will partially undermine the American democratic system.

In this sense, these measures partially undermine the democratic system of the United States.



BLEU = 26.52, 75.0/40.0/21.4/7.7 (BP=1.000, ratio=1.143, hyp_len=16, ref_len=14)

$$\log \text{BLEU} = \min(1 - \frac{r}{c}, 0) + \sum_{n=1}^{N} w_n \log p_n.$$

In our baseline, we use $N = 4$ and uniform weights $w_n = 1/N$.

Papineni, Kishore, et al. "BLEU: a method for automatic evaluation of machine translation." Proceedings of the 40th annual meeting on association for computational linguistics. Association for Computational Linguistics, 2002.

Statistical Machine Translation (2002 - 2015)

Data-Driven Machine Translation

Target language corpus gives examples of well-formed sentences

I will get to it later

See you later

He will do it

Parallel corpus gives translation examples

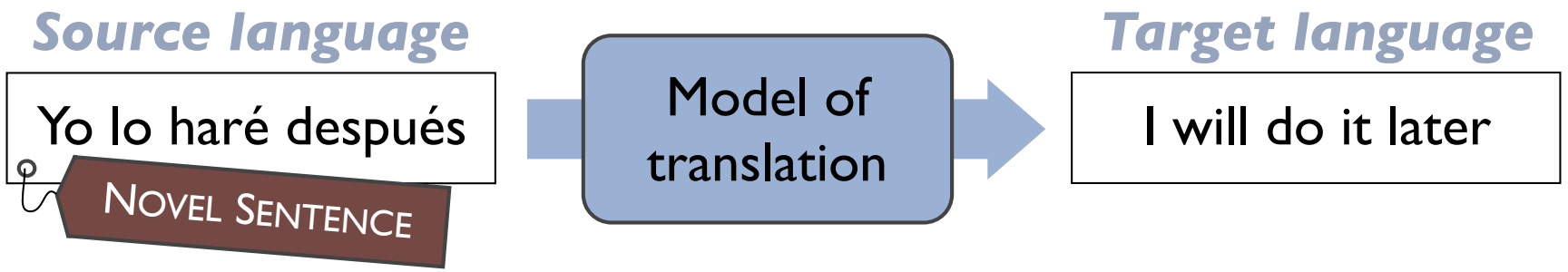
I will do it gladly

Yo lo haré de muy buen grado

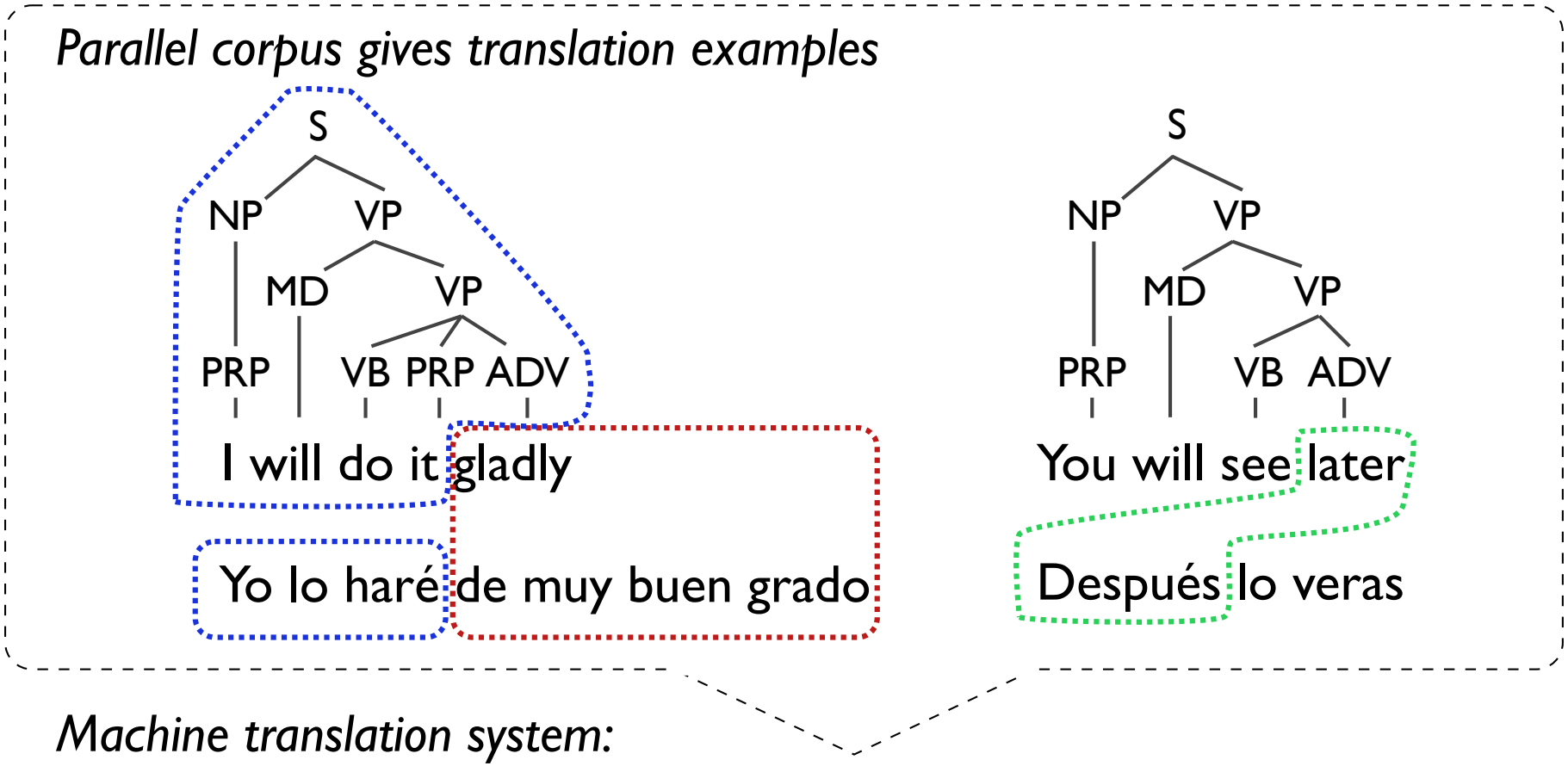
You will see later

Después lo veras

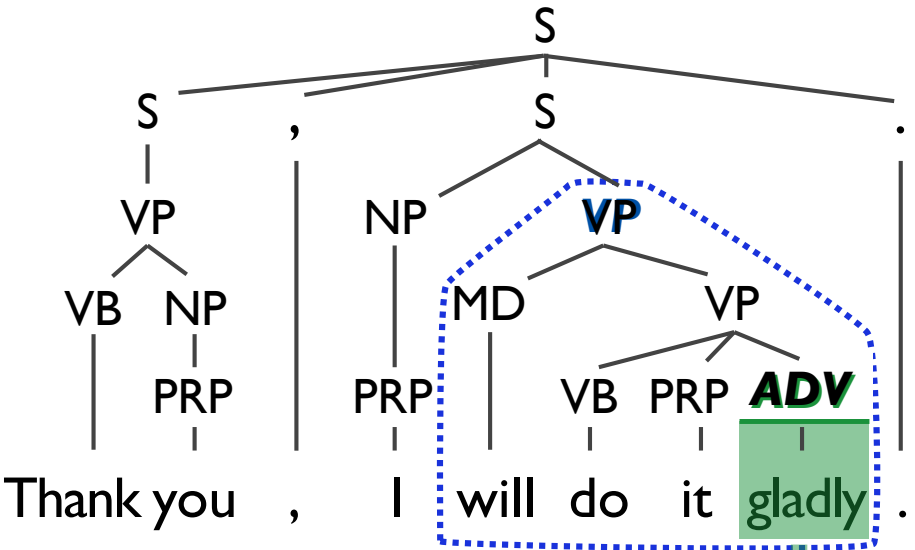
Machine translation system:



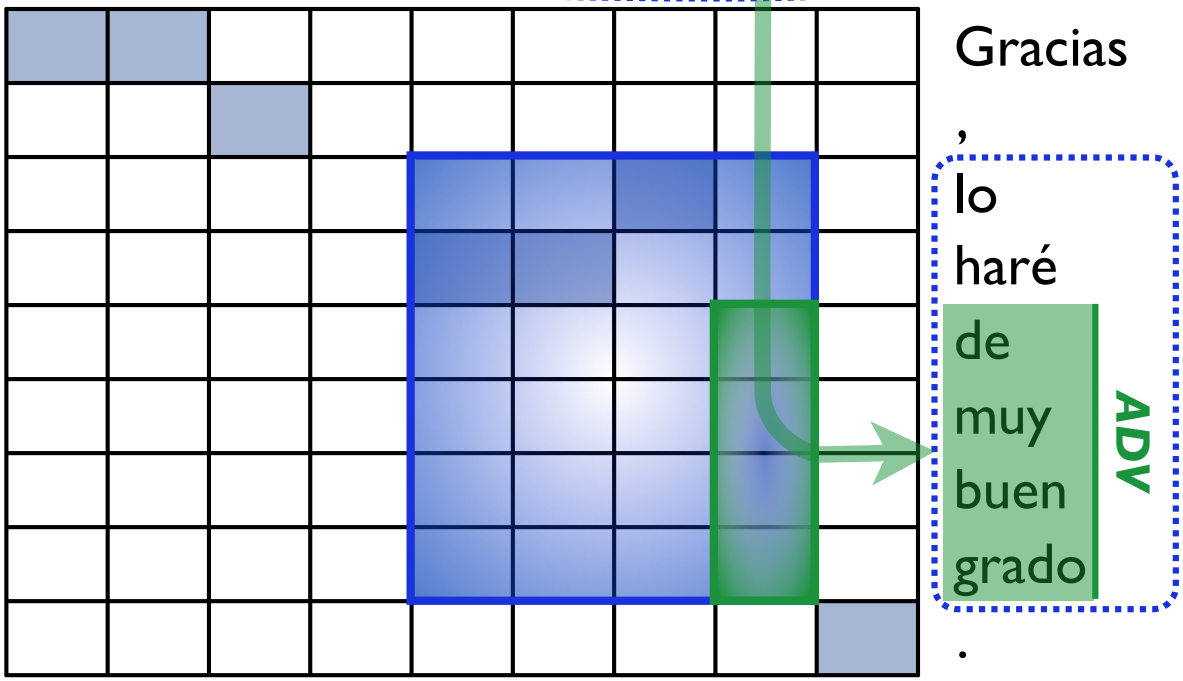
Stitching Together Fragments

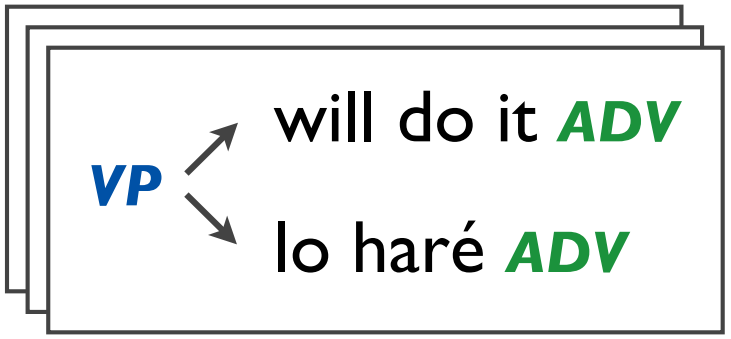


Extracting Translation Rules

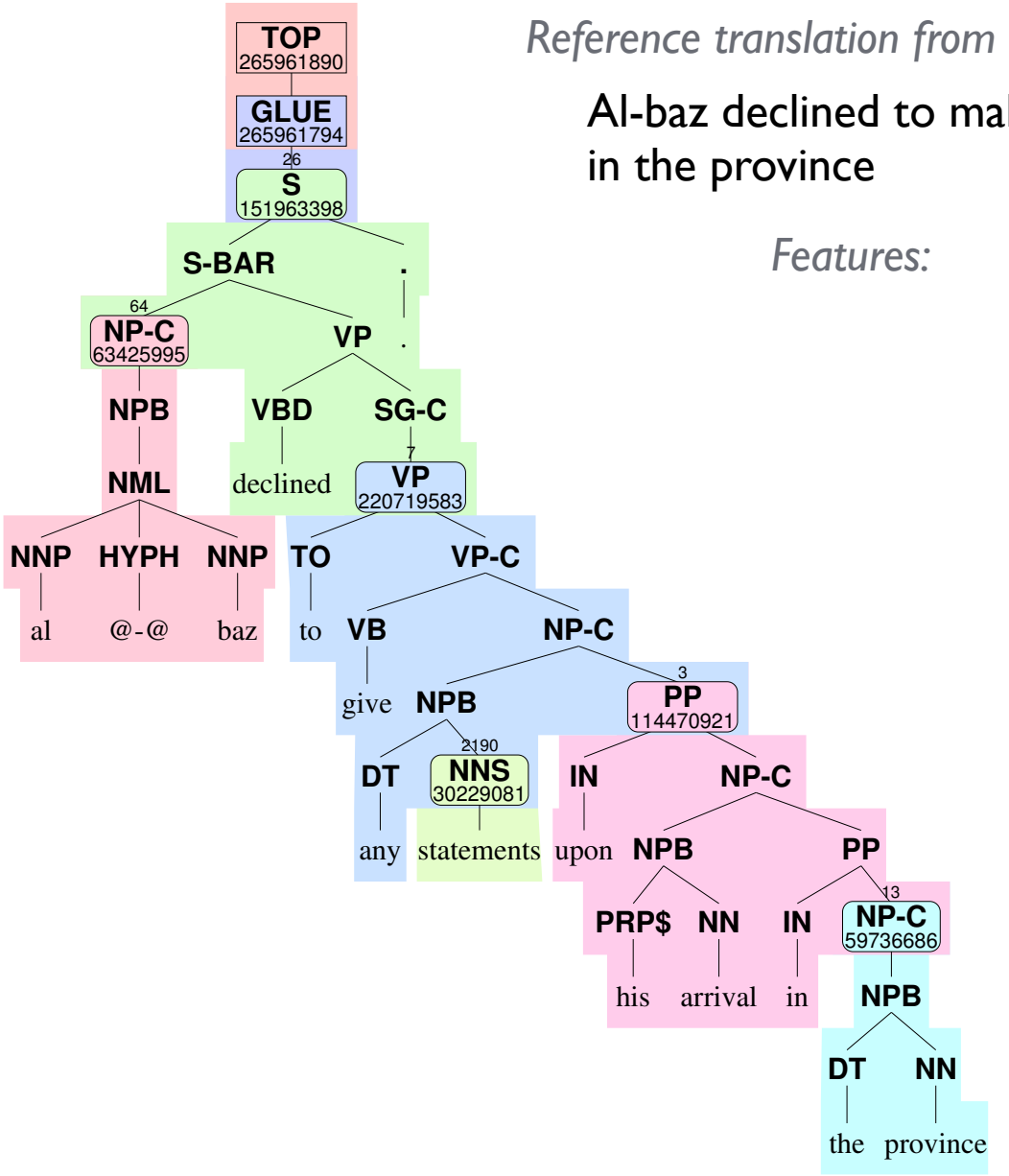


Frequency statistics on these rules guide translation





An Example Syntax-Based Translation



Reference translation from a human translator:

Al-baz declined to make any statements upon his arrival in the province

Features:

feature	weight	value	product
derivation-size	0.41	8	3.30
glue-rule	3.89	2	7.78
green	-0.08	0	0
gt_prob	0.40	36.18	14.43
identity	-9.97	0	0
is_lexicalized	-0.65	6	-3.91
lex_pef	1.02	5.47	5.60
lex_pfe	0.31	4.44	1.39
lm1	1	22.76	22.76
lm1-unk	30.08	0	0
lm2	0.74	26.66	19.79
lm2-unk	-39.18	0	0
missingWord	-1.29	0	0
model1inv	1.02	10.60	10.81
model1nrm	1.35	11.29	15.22
nonmonotone	4.17	0	0
olive	1.95	0	0
psm1n	0.50	24.65	12.30
text-length	-3.87	15	-58.05
trivial_cond_prob	0.41	3.34	1.38
unk-rule	19.28	0	0
reported **totalcost**	52.82	$\vec{v} \cdot \vec{w}$	52.82

Discrete Word Alignment Models

Classical independence assumption (IBM Model 2; Brown et al. '93):

$$P(\mathbf{e}|\mathbf{f}) = \prod_{i=1}^{|\mathbf{e}|} P(e_i \mid f_{a_i}) \cdot P(a_i = j \mid |\mathbf{e}|, |\mathbf{f}|)$$

	set	series	whole
série	0.1	0.7	0.1
ensemble	0.6	0.1	0.2
set	0.9	0.0	0.0
fixé	0.6	0.0	0.0

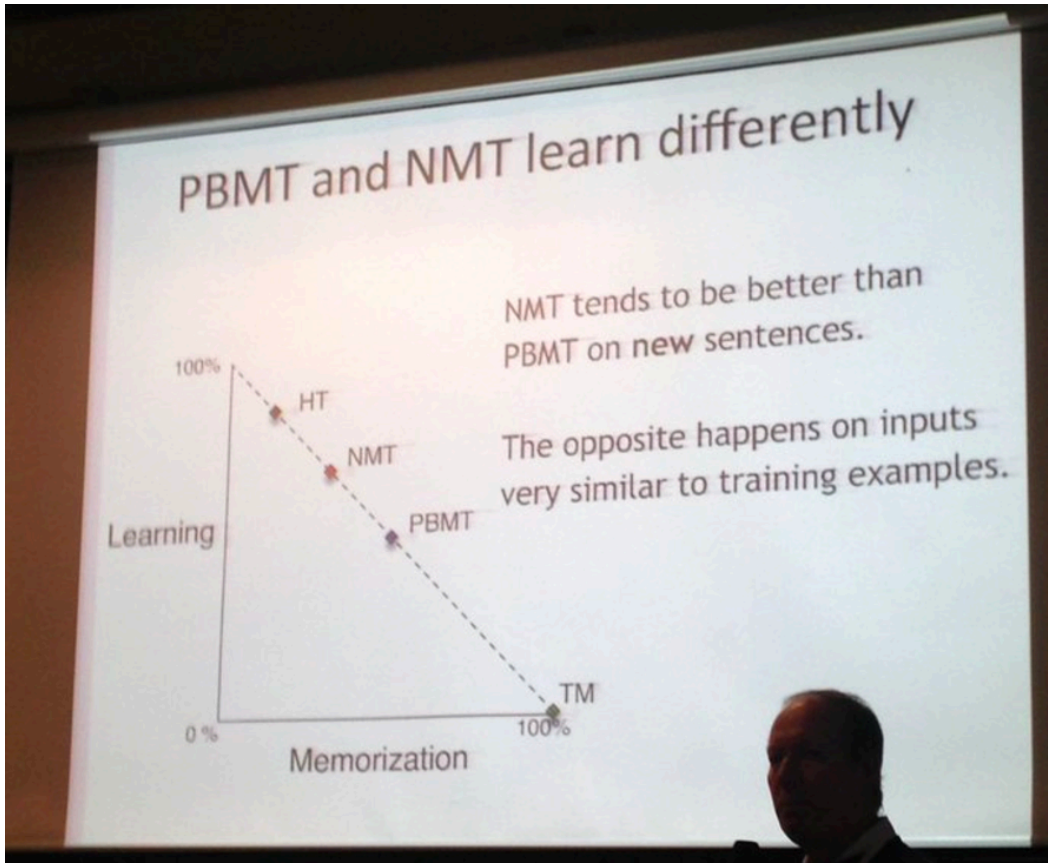
Correspondence between word positions depends only on sentence lengths, not contents

Brown, Peter F., et al. "The mathematics of statistical machine translation: Parameter estimation." *Computational linguistics*19.2 (1993): 263-311.

Structured Translation Rules

Phrase-based: "Upon his arrival in" <-> "a su llegada a"

Syntax-based: "(PP upon PRP$$_1$$ arrival in)" <-> "a PRP$$_1$$ llegada a"



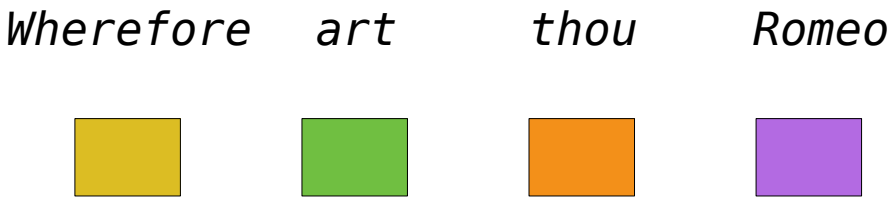
Marcello Federico (FBK), 2017

Idioms are idiosyncratic phrases such as, "the benefit of the doubt"

Neural Machine Translation (2016-present)

Word Embeddings

Each discrete element of text (word, subword, or character) has a dense vector, called its embedding:



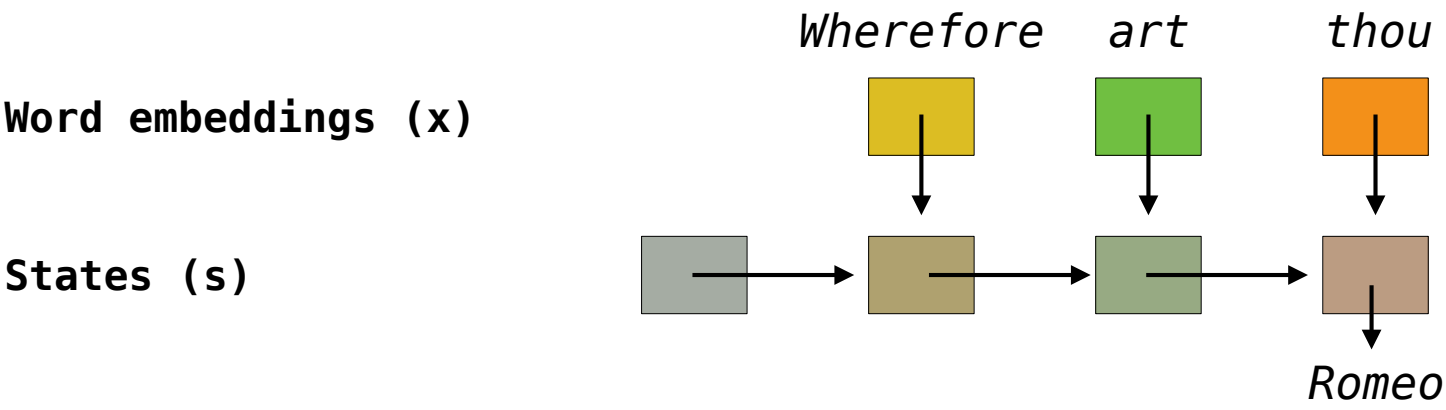
An embedding matrix is |V| x k; typical |V| is 50,000 & k is 1,000.

Most of the parameters in a neural MT model are in the two embedding matrices, one for input and one for output.

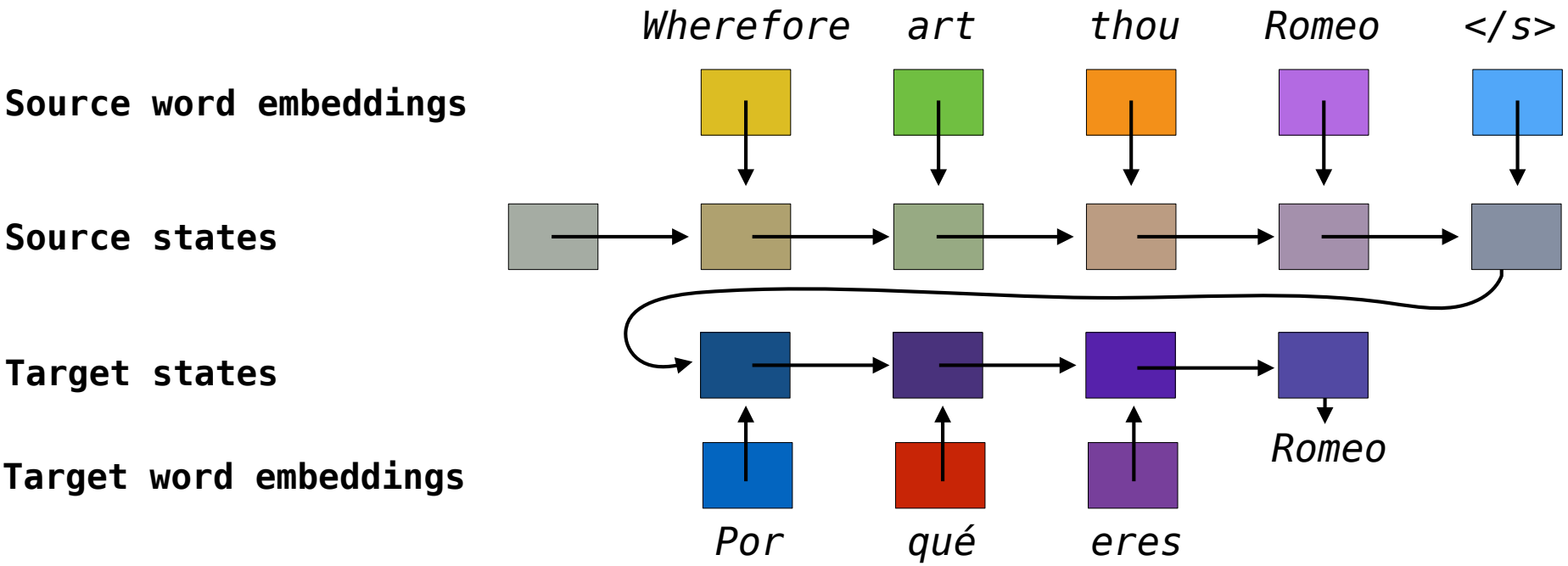
Embeddings are learned by optimizing a complex translation model; Initialization is largely irrelevant for typical data conditions.

Encoder-Decoder Framework

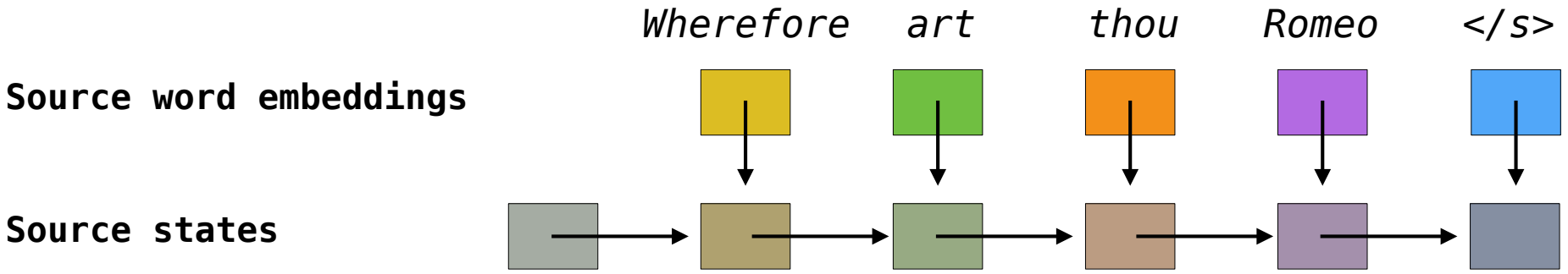
Language modeling: predict a word given its left context.



Translation: predict a word given its left context & the source.



Recurrent Units



The simplest type of "recurrent neural network" doesn't work well for full sentence representation:

$$s_t = g(x_t W^x + s_{t-1} W^s + b^s)$$

Good idea #1: Compute dimension i of s_t from dimension i of s_{t-1}.
Good idea #2: Let the input "reset" dimension i of s_t as needed.
Long short-term memory (LSTM), **gated recurrent unit (GRU)**, etc.

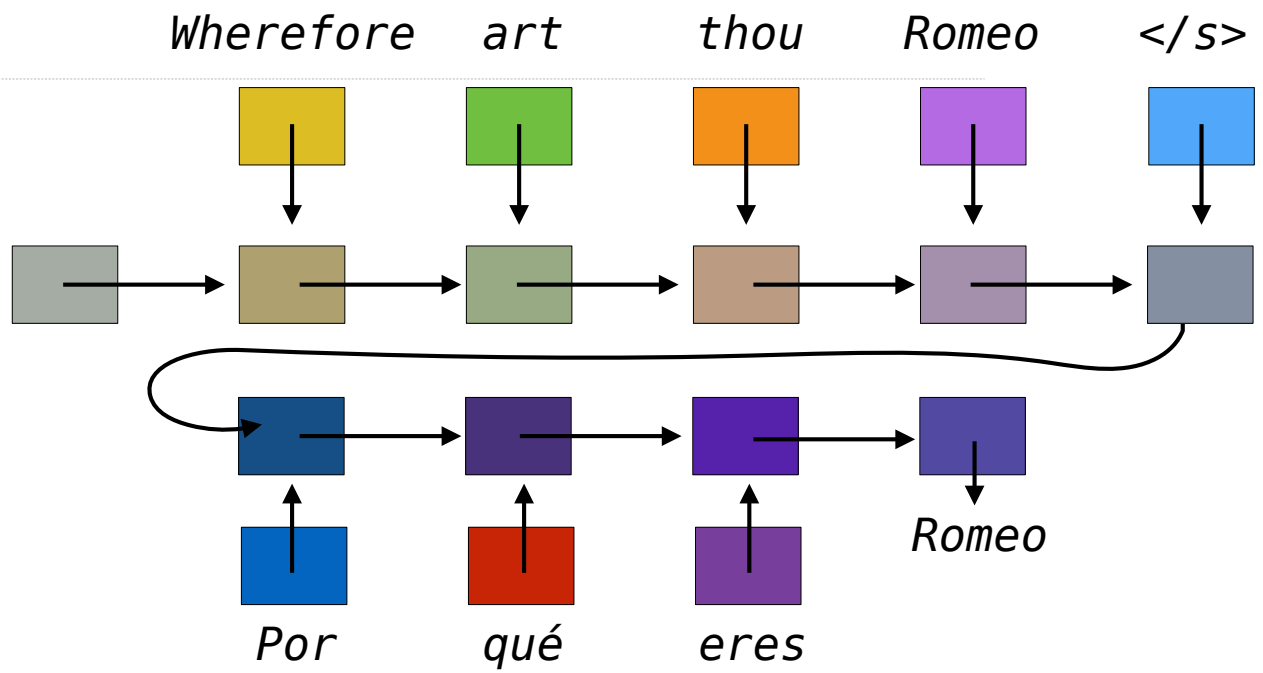
$$s_t = z_t \odot s_{t-1} + (1 - z_t) \odot g(x_t W^x + (r_t \odot s_{t-1}) W^s + b^s)$$

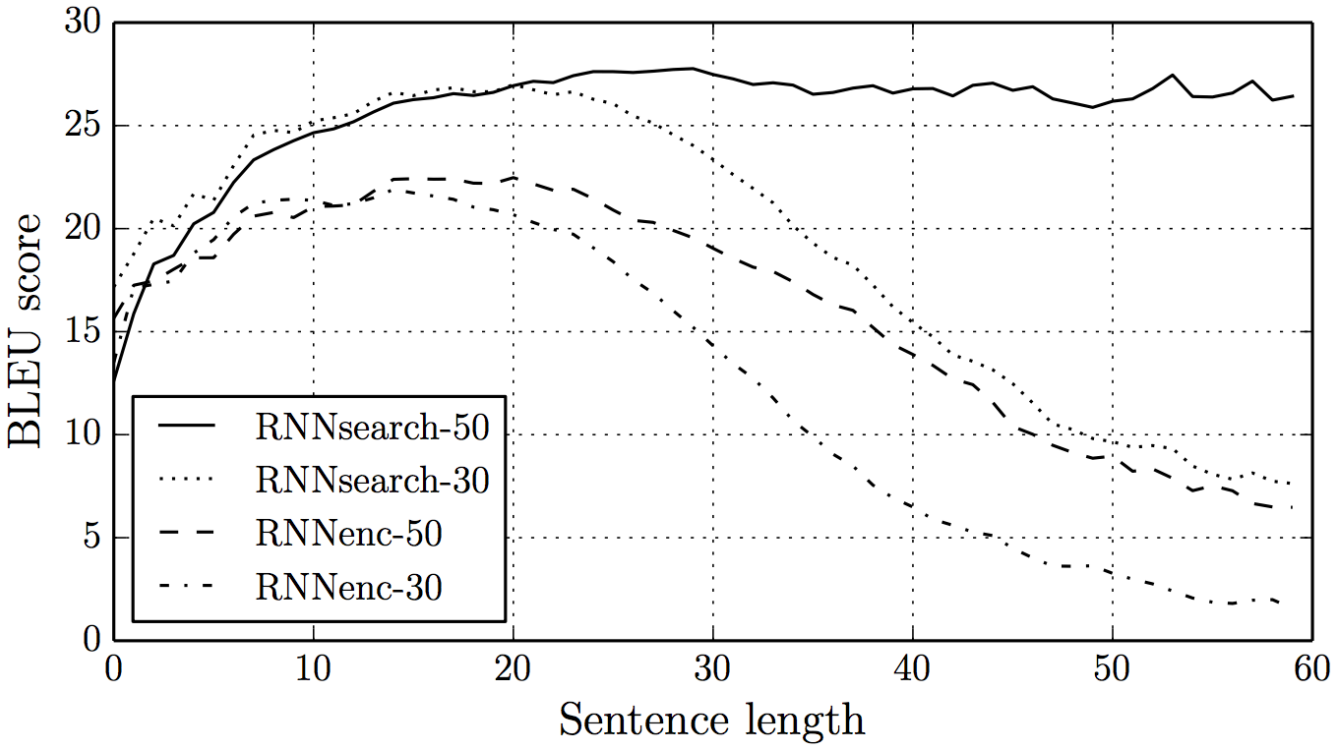
$$z_t = f_z(x_t, s_{t-1})$$

$$r_t = f_r(x_t, s_{t-1})$$

Vectors that determine what happens to each dimension of s_{t-1}.

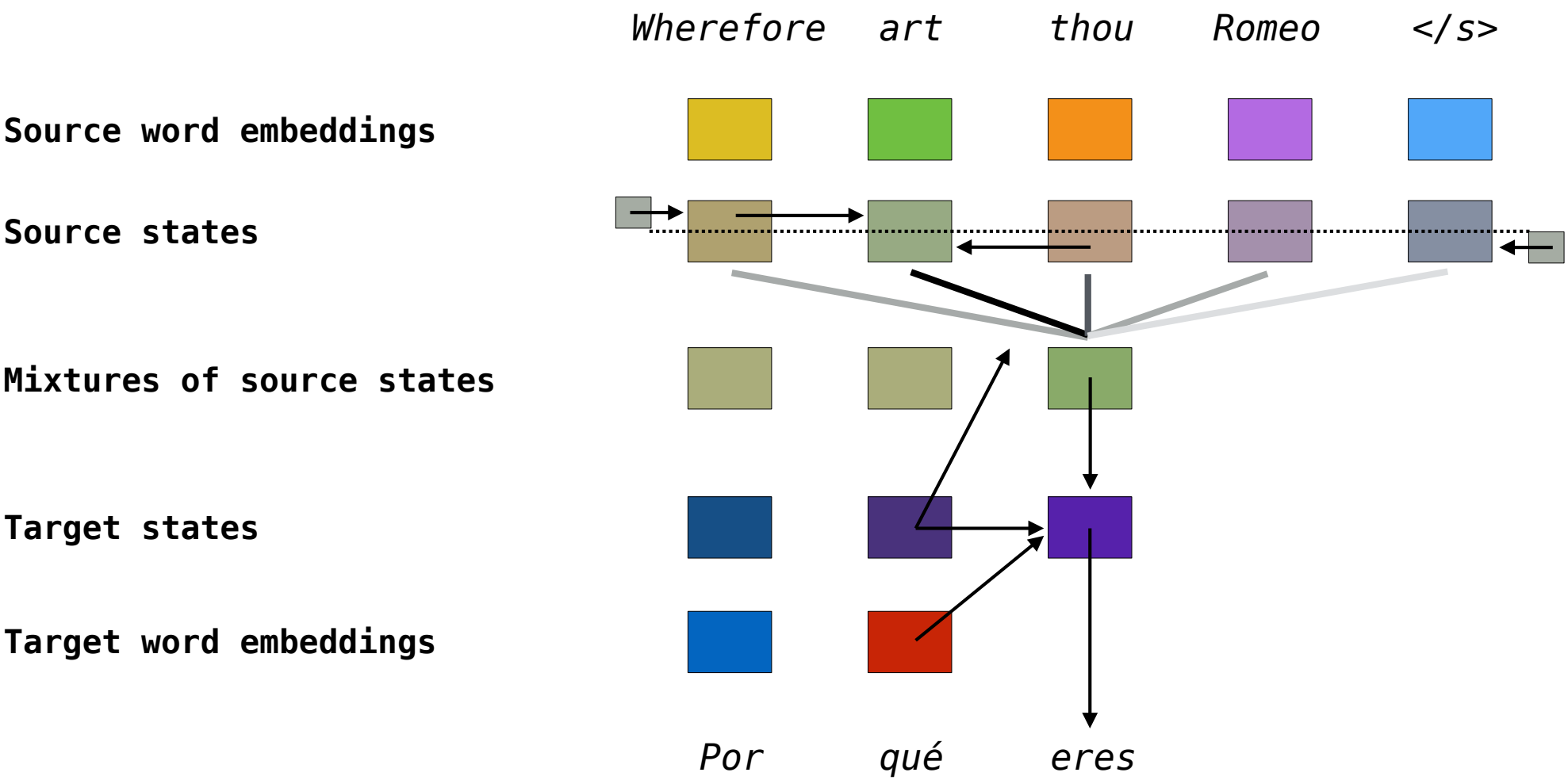
Attention





Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. "Neural machine translation by jointly learning to align and translate." arXiv preprint arXiv:1409.0473 (2014).

Neural Machine Translation with Attention



Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. "Neural machine translation by jointly learning to align and translate." arXiv preprint arXiv:1409.0473 (2014).

System Comparison

Recent English->German Results

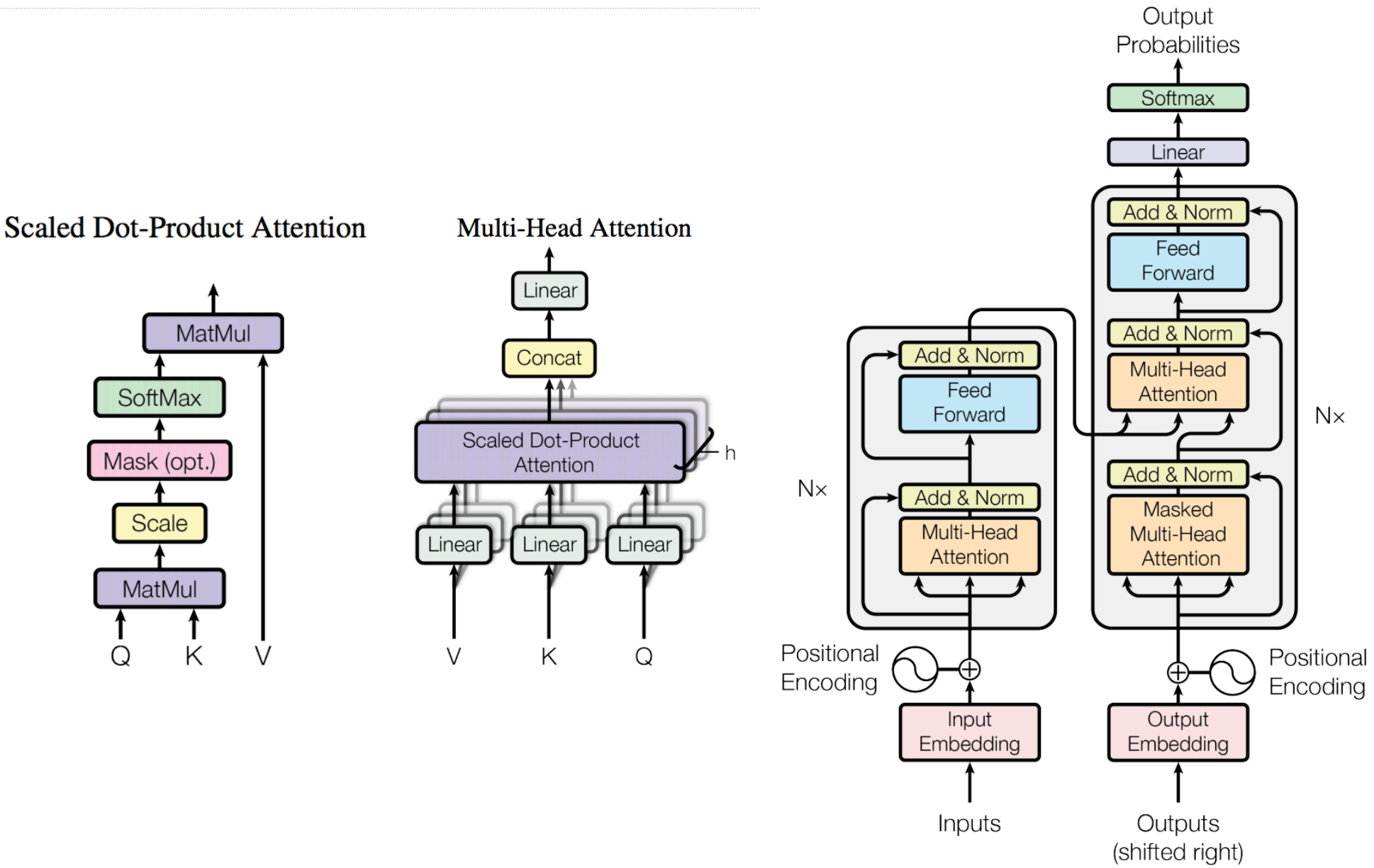
System Description	BLEU newstest2014
Statistical MT (Sennrich & Haddow, 2015)	22.6
Neural MT with Attention (Bahdanau et al., 2014)	19.9
+ Synthetic training data (Sennrich et al., 2016)	22.7
+ Ensemble of neural models (Sennrich et al., 2016)	23.8
+ Deep network & RL objective (Wu et al., 2016)	26.3
Transformer network (Vaswani et al., 2017)	28.4

[Sennrich et al. *Improving Neural Machine Translation Models with Monolingual Data.*]
[Bojar et al. *Findings of the 2016 Conference on Machine Translation.*]
[Wu et al. Google's Neural Machine Translation System.]
[Vaswani et al. Attention is All You Need.]

The Transformer

[Vaswani et al. Attention is All You Need. NIPS 2017]

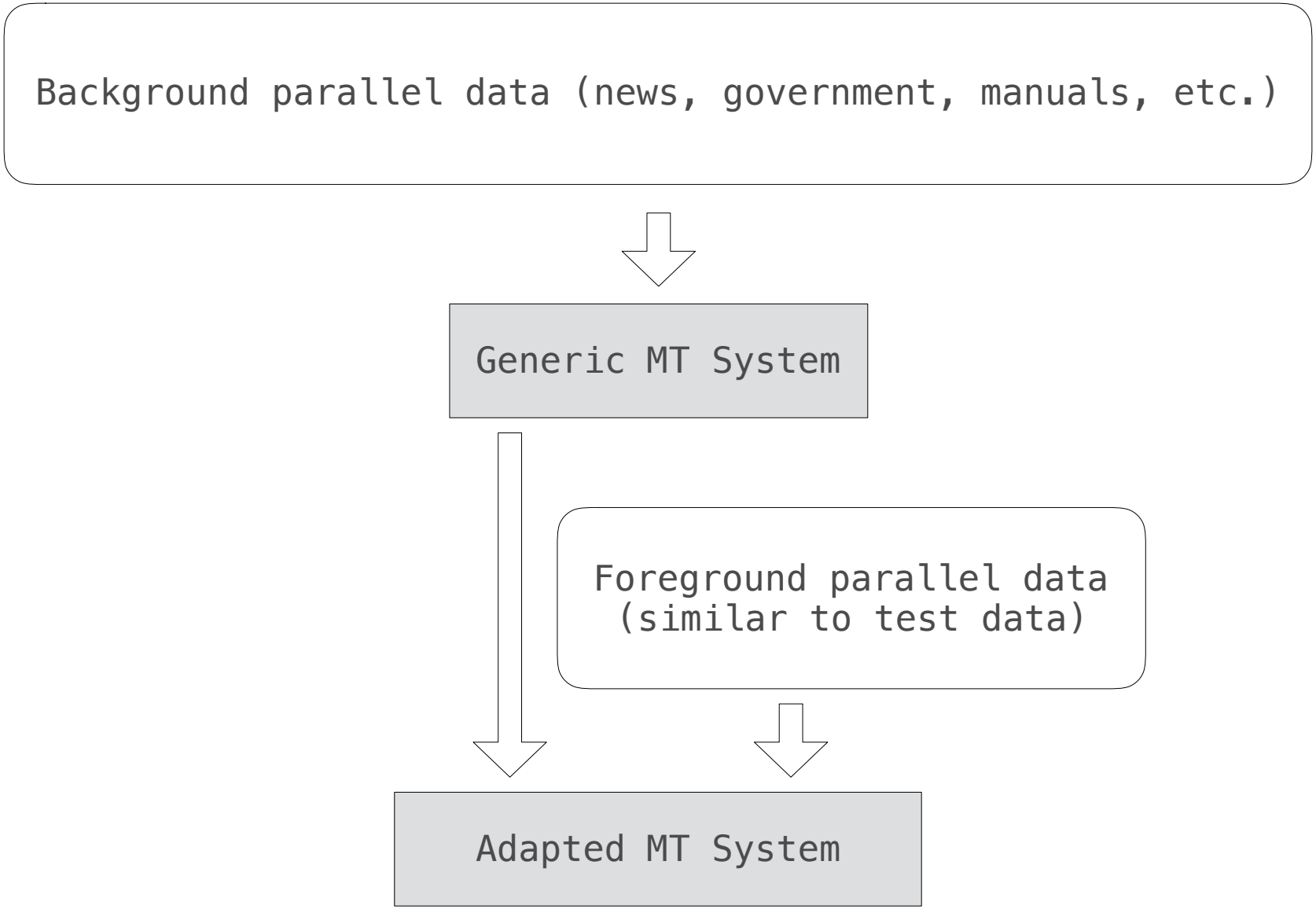
The Transformer



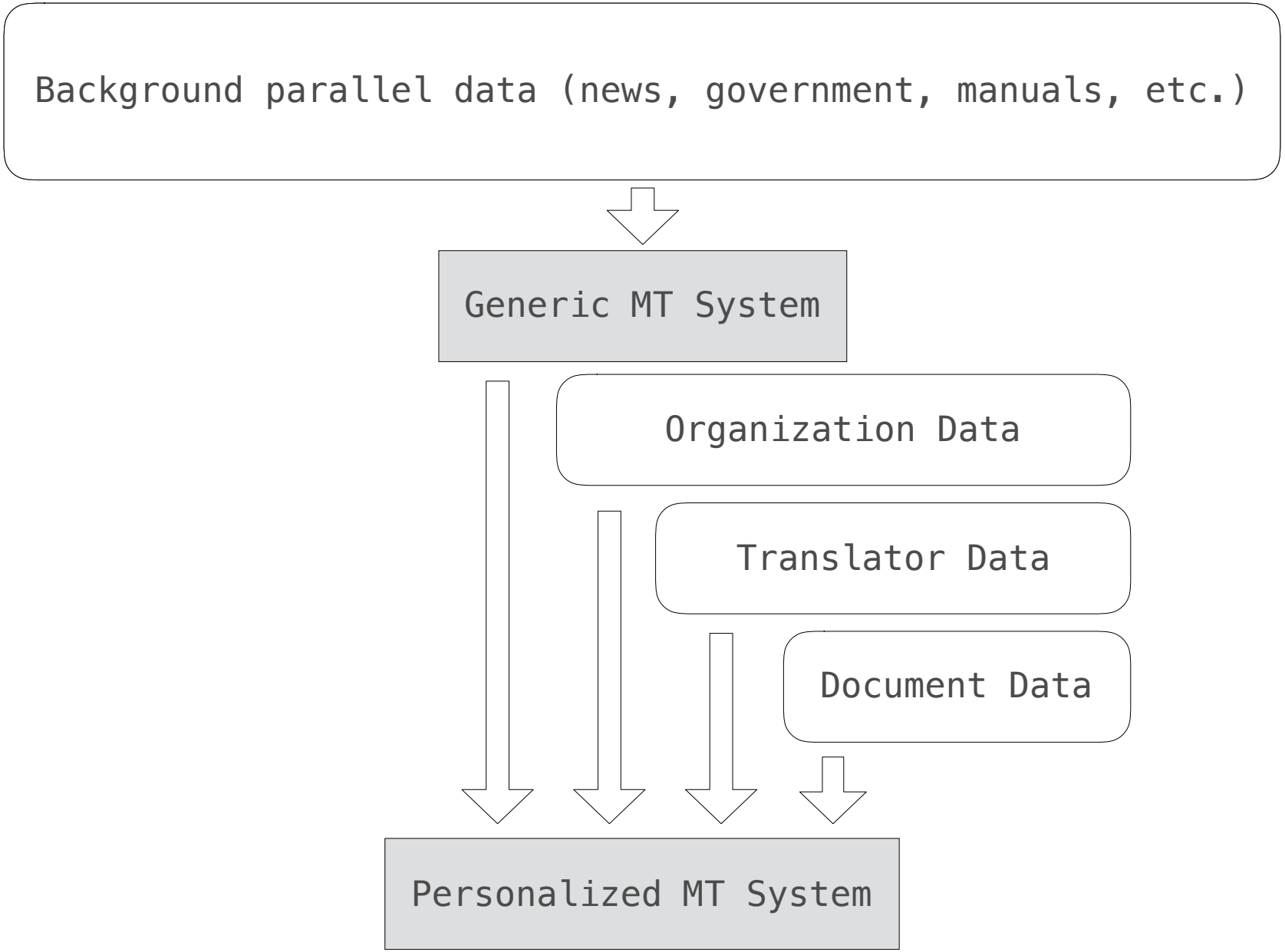
[Vaswani et al. Attention is All You Need. NIPS 2017]

Adaptive Machine Translation

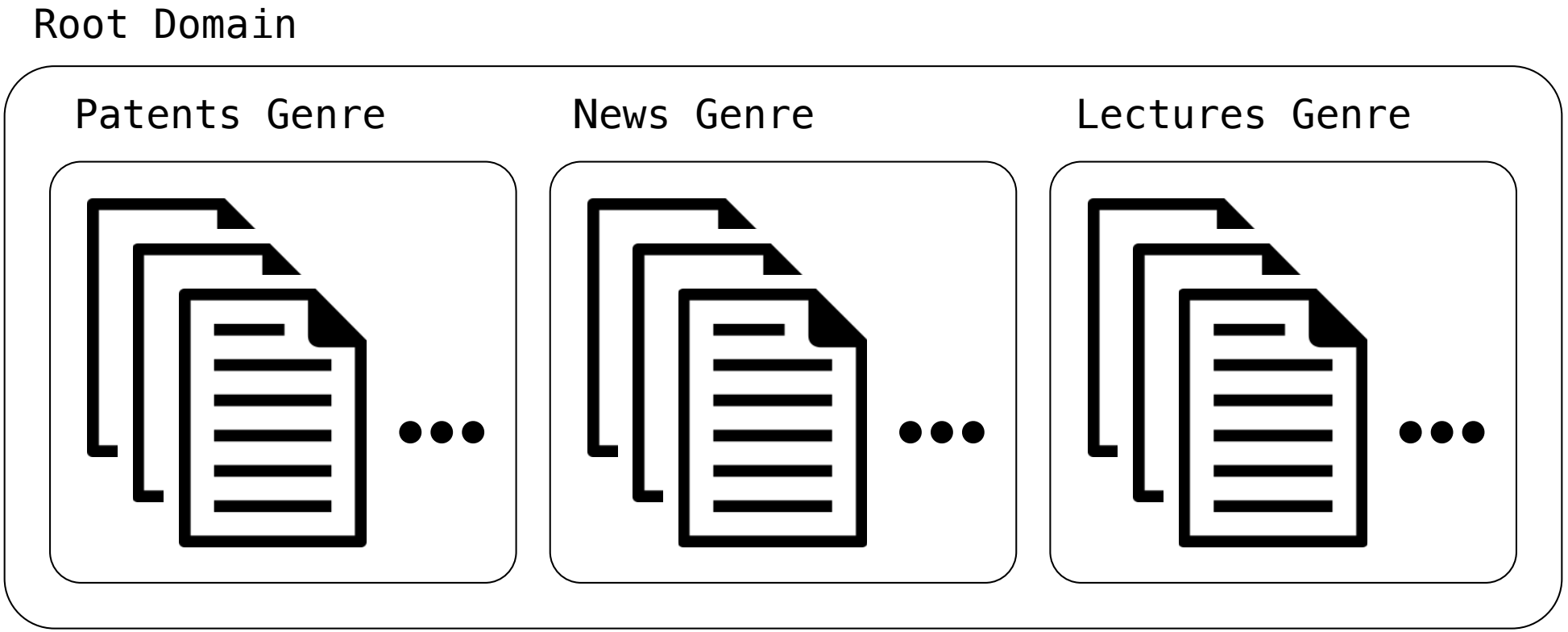
Standard Model Adaptation



Hierarchical Model Personalization



Hierarchical Model Personalization for Statistical MT



$$w_{ROOT} \cdot f_{ROOT} + w_{GENRE} \cdot f_{GENRE} + w_{DOC} \cdot f_{DOC}$$

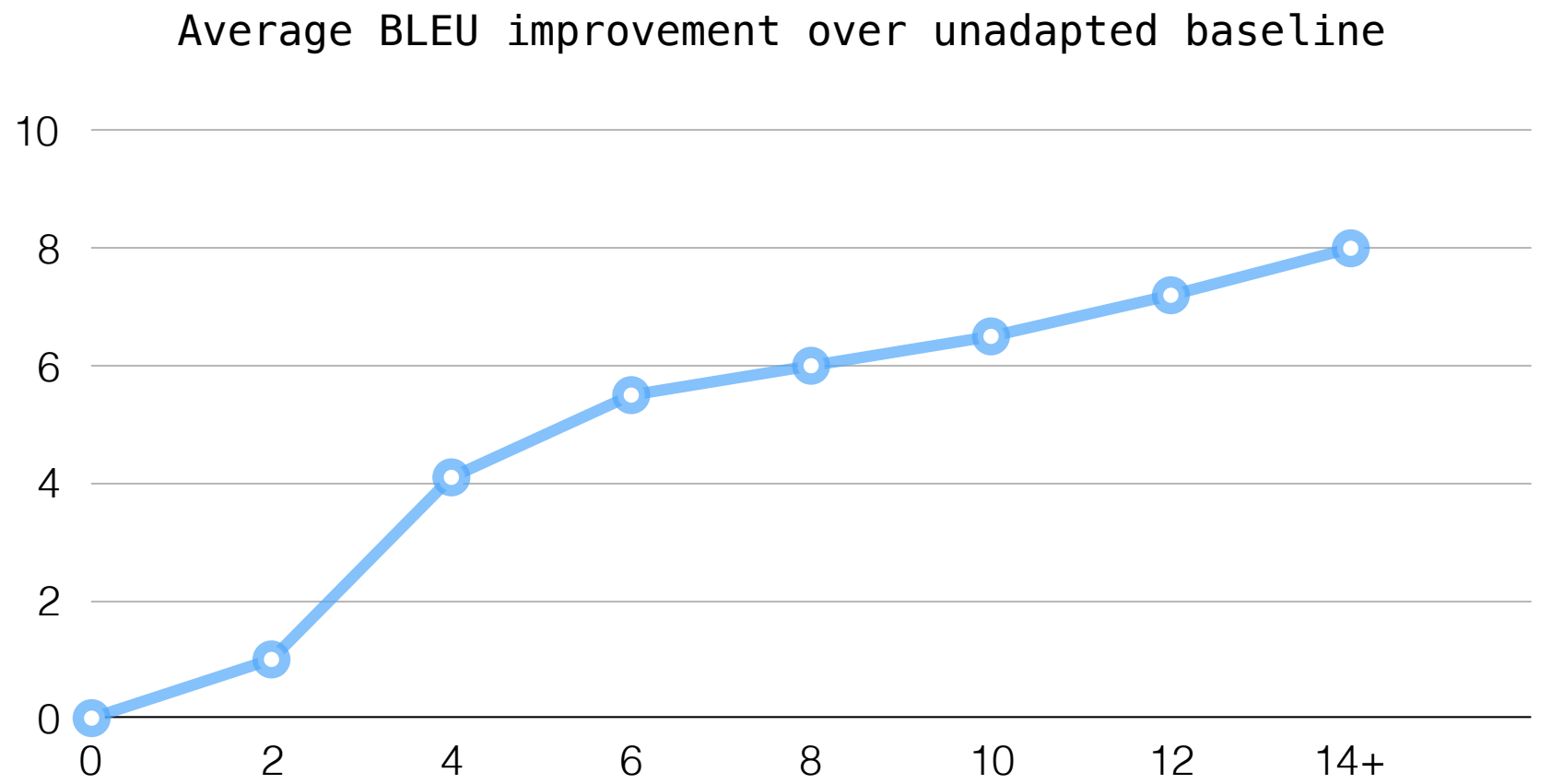
Learned in batch

Estimated from full corpus

Learned online

Estimated from domain corpus

Hierarchical Adaptation Results for Patent Translation



Overall, +5.3 BLEU in German–English patent translation in SMT

[Wuebker et al. Hierarchical Incremental Adaptation for Statistical Machine Translation]

Adapting a Neural Model

Basic adaptation is much more effective for neural MT systems (unpublished)

	Baseline	Adapted
Statistical MT	27.1	34.7 (+7.6)
Neural MT	27.4	38.4 (+11.0)

Interactive Machine Translation

Interactive & Adaptive Machine Translation

(Demo)

with Spence Green, Joern Wuebker, & Sasa Hasan

Technical problems:

- Online learning for user adaptation [EMNLP '15]
- Translation inference for suggestions [ACL '16]

Production constraints:

- Latency below 500ms
- Improvements should be immediately perceptible

Auto-Complete Results

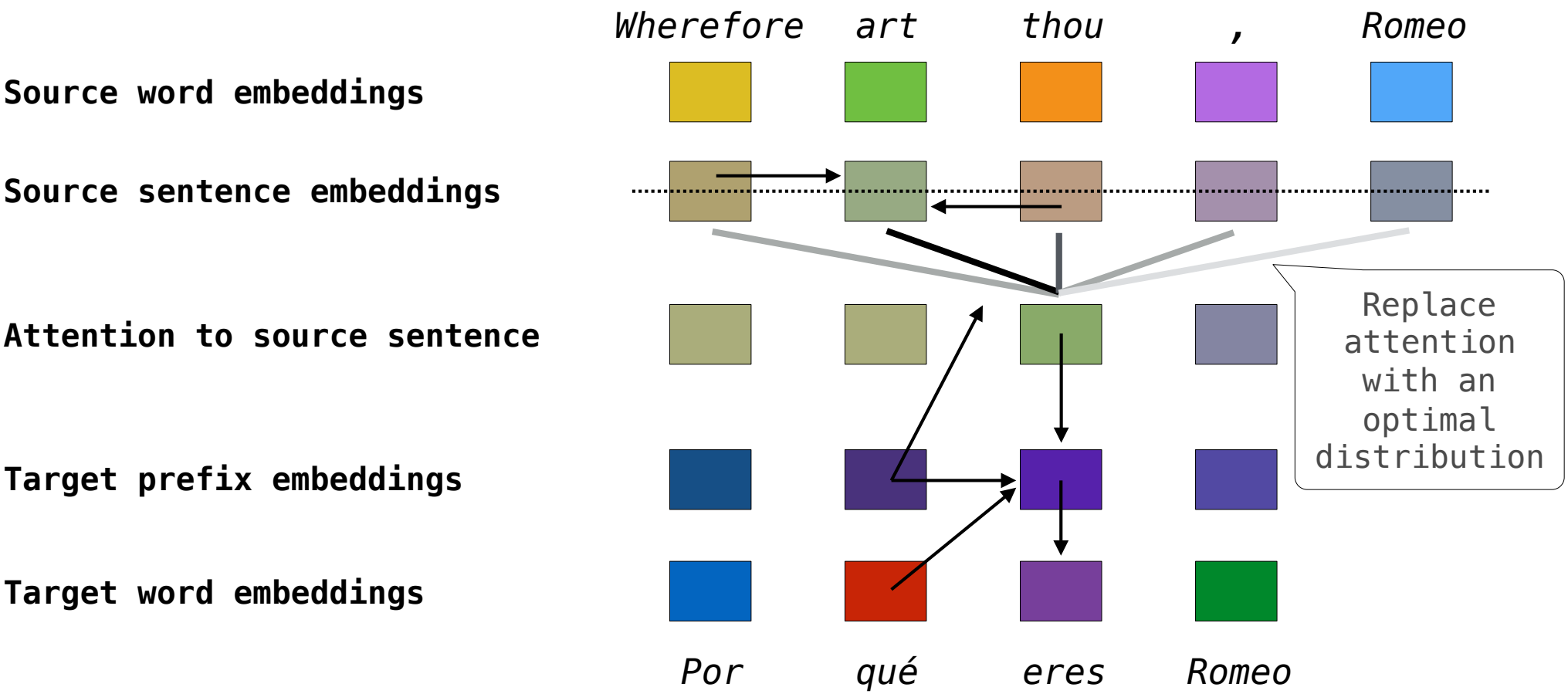
English–German Software Domain (ACL, 2016)

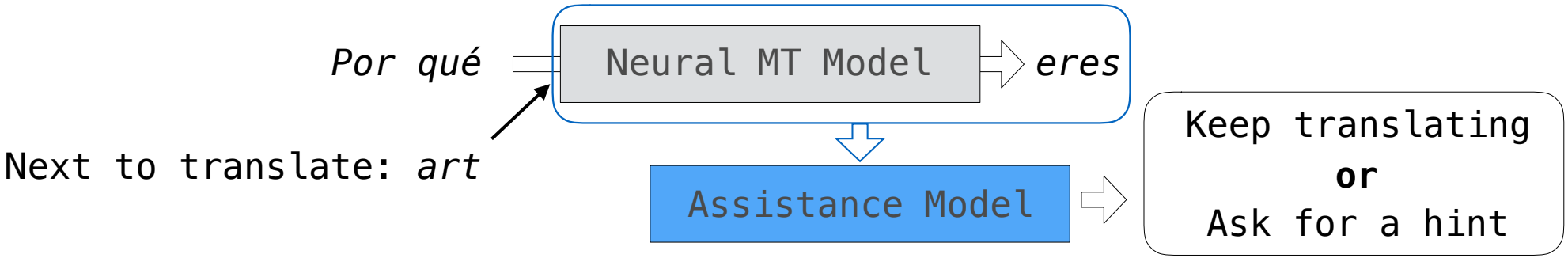
	BLEU of whole sentences	BLEU of suffix	Accuracy of first word of suffix
Statistical MT baseline	44.5	58.8	37.8
+ Prefix improvements	44.5	62.2	46.0
Neural MT	44.3	64.7	54.9

English–German News (newstest2015; unpublished, 2017)

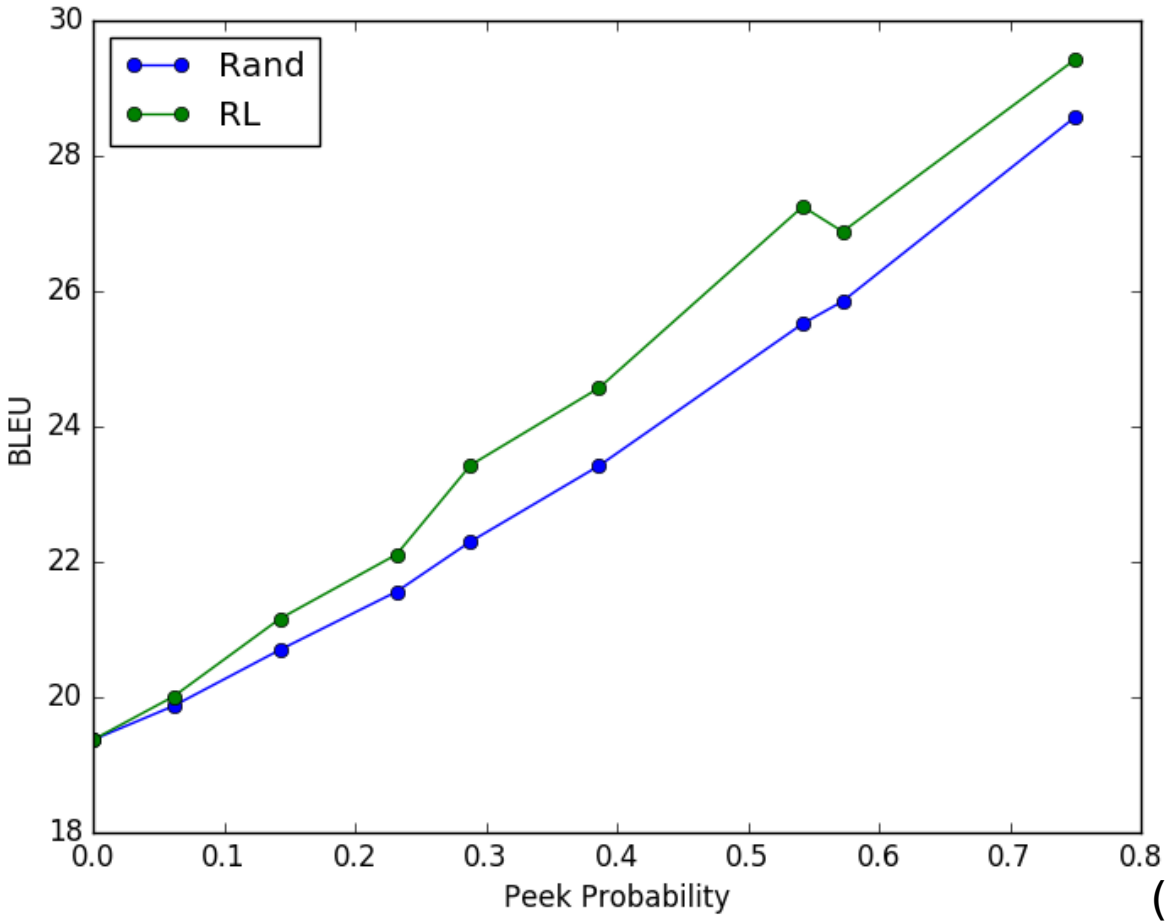
Statistical MT	22.9	36.1	40.6
Neural MT	24.0	41.7	54.3

Interactive Attention for Neural Translation





Learning an Interactive Attention Policy for Neural Machine Translation



(MT Summit, 2017)