Why Neural Translations are the Right Length

Xing Shi, Kevin Knight and Deniz Yuret;

EMNLP 2016

Information Sciences Institute
USC Viterbi School of Engineering

What is the fundamental question as a **PhD student** ?

How to publish a lot of high-quality papers ?

How to graduate in 5 years ?

PhD Life

MT

How to publish a lot of high-quality papers ?

How to graduate in 5 years ?

PhD Life	MT
How to publish a lot of high-quality papers ?	H-index || BLEU
How to graduate in 5 years ?	5 years || right length

Language Pairs	BLEU	Length Ratio (MT output / reference)
English => Spanish	31.0	**0.97**
English => French	29.8	**0.96**

2-layer 1000 hidden units non-attentional LSTM seq2seq

English :	does he know about phone hacking ?
French reference :	a-t-il connaissance du piratage téléphonique ?
French translation:	<UNK> <UNK> <UNK> <UNK> ?

	When to stop
PBMT	[- - - -] → [- x - -] → [x x x x]
Neural MT	Word → Word → <EOF>

	When to stop	How to generate right length ?
PBMT	[- - - -] → [- x - -] → [x x x x]	● word-penalty feature
Neural MT	Word → Word → <EOF>	● no explicit penalty

	When to stop	How to generate right length ?
Statistical MT	[- - - -] → [- x - -] → [x x x x]	• word-penalty feature • MERT
Neural MT	Word → Word → <EOF>	• no explicit penalty • MLE

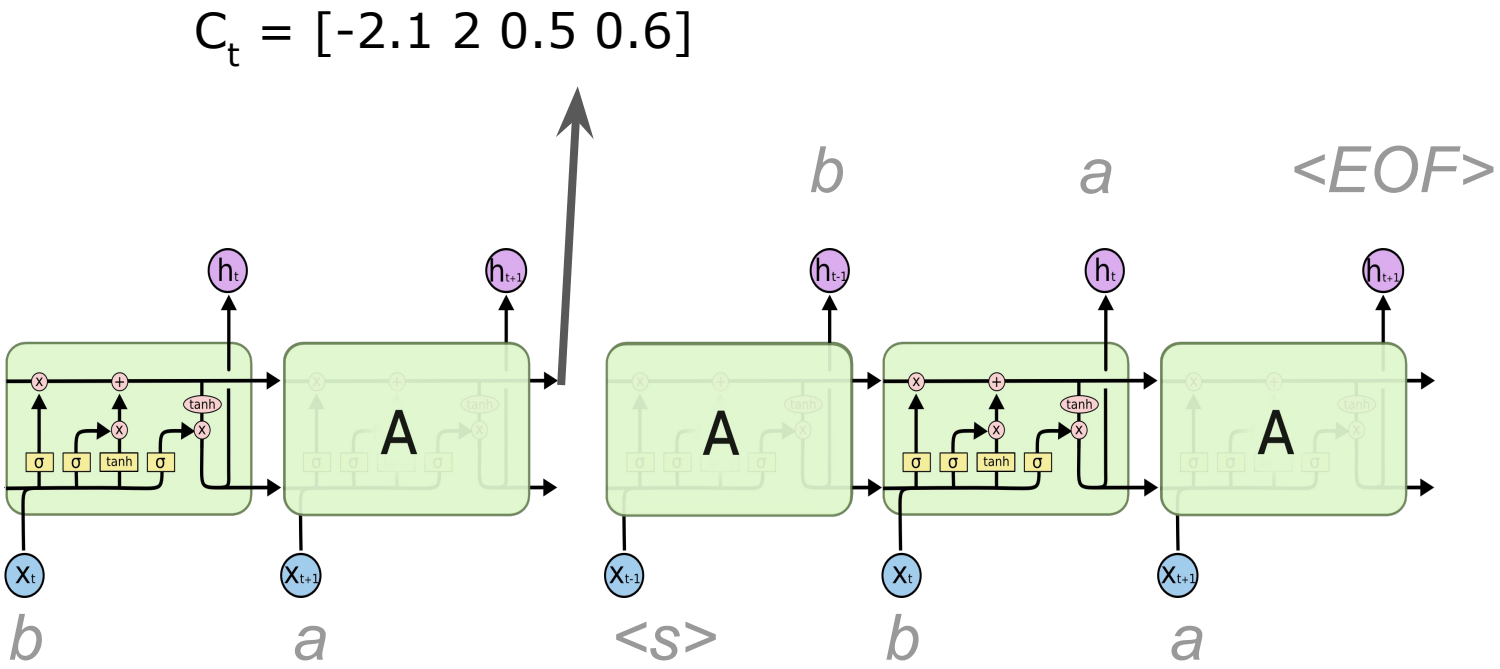
	When to stop	How to generate right length ?
Statistical MT	[- - - -] → [- x - -] → [x x x x]	● word-penalty feature ● MERT ● Heavy beam search
Neural MT	Word → Word → <EOF>	● no explicit penalty ● MLE ● light beam search (beam = 10)

Toy Example: String Copy

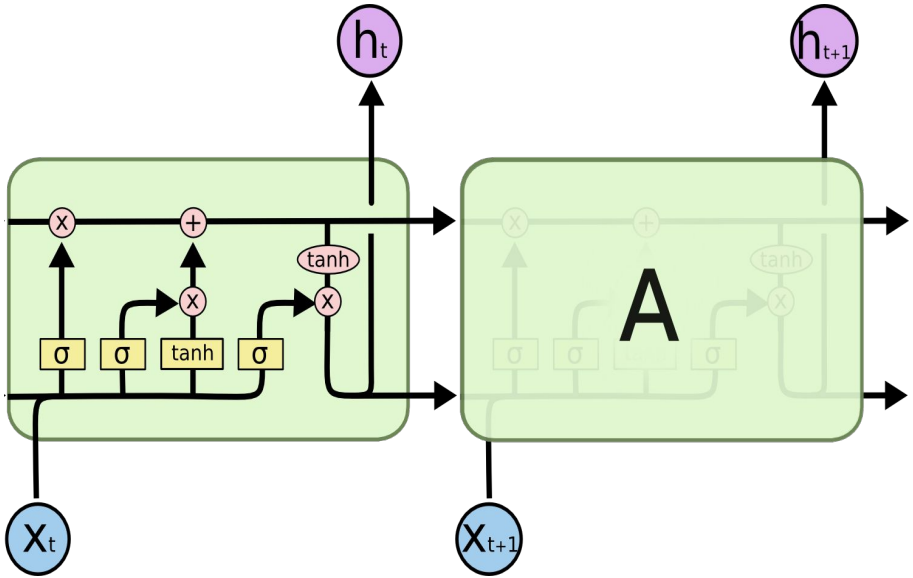
a a a b b <EOS> → *a a a b b <EOS>*
b b a <EOS> → *b b a <EOS>*

Train: 2500 random string
Single-layer, 4 hidden states LSTM

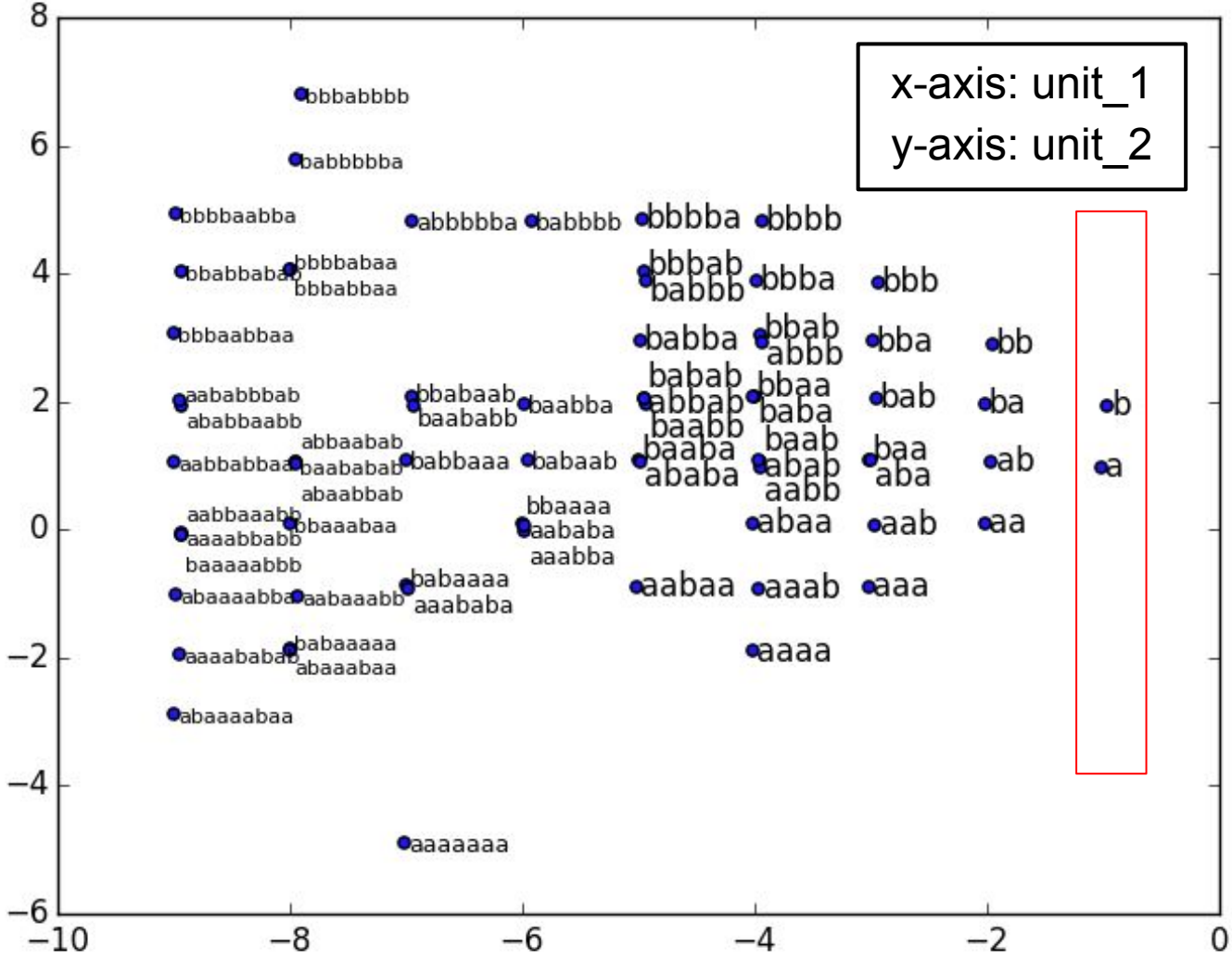
Toy Example: String Copy

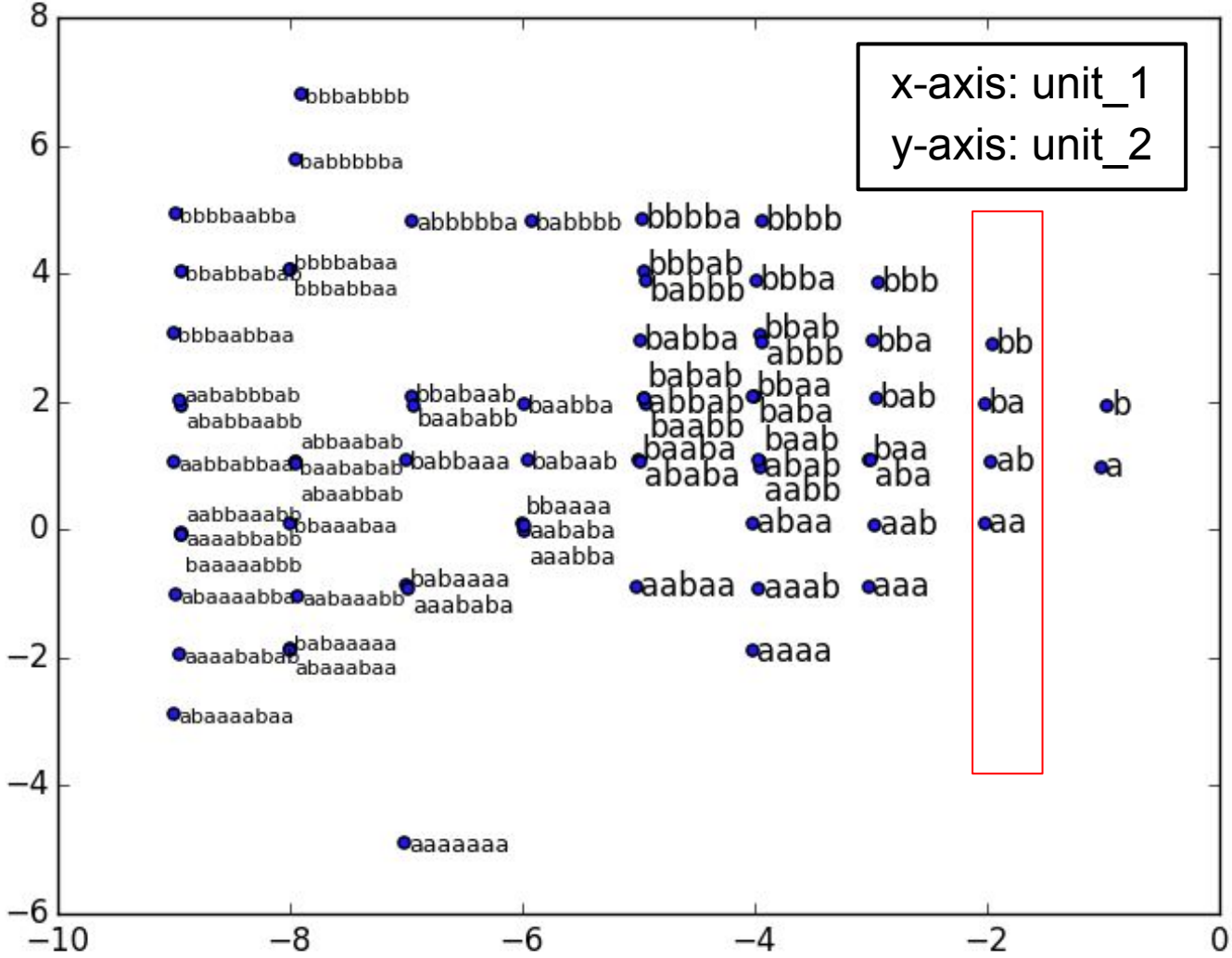


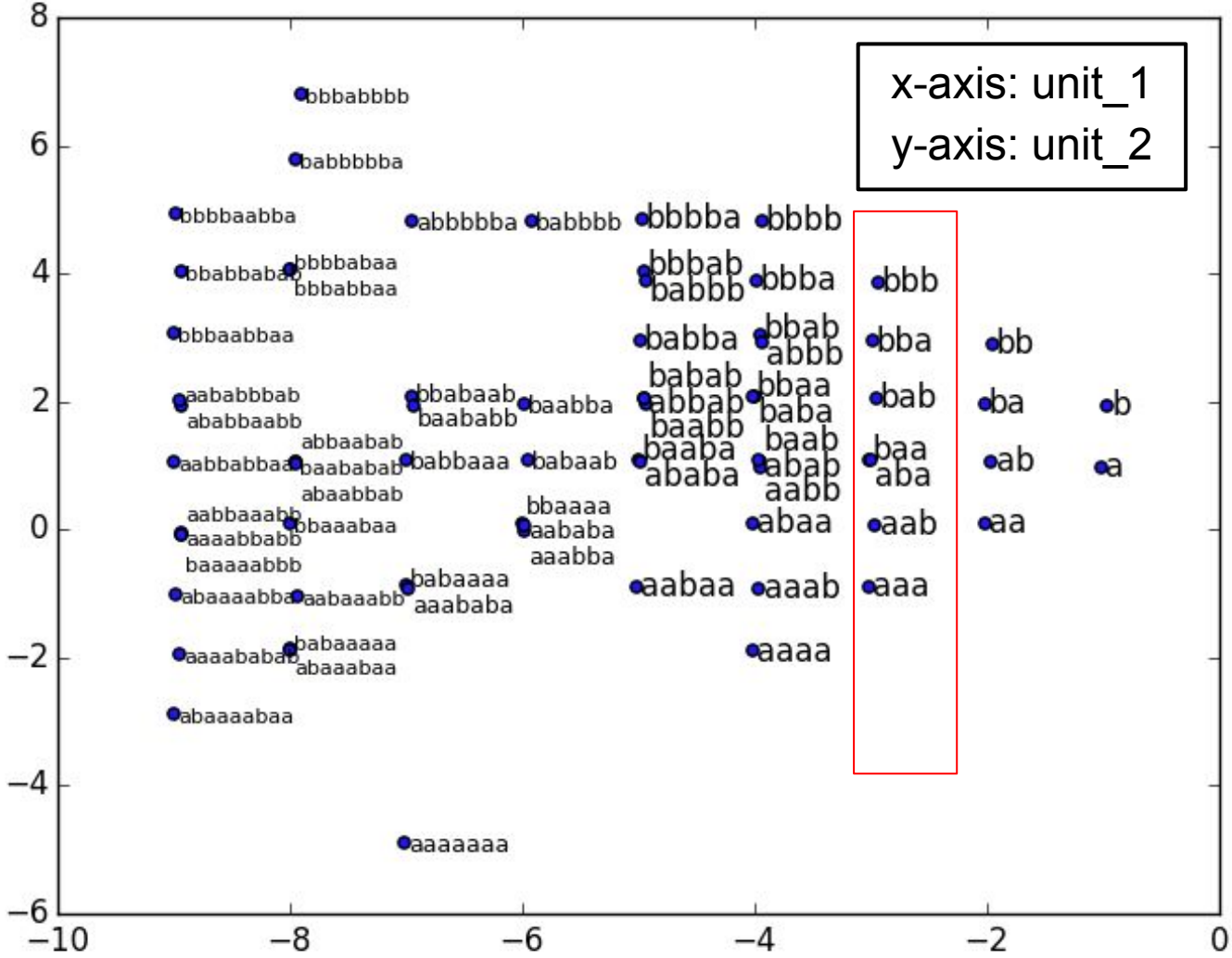
Toy Example: String Copy

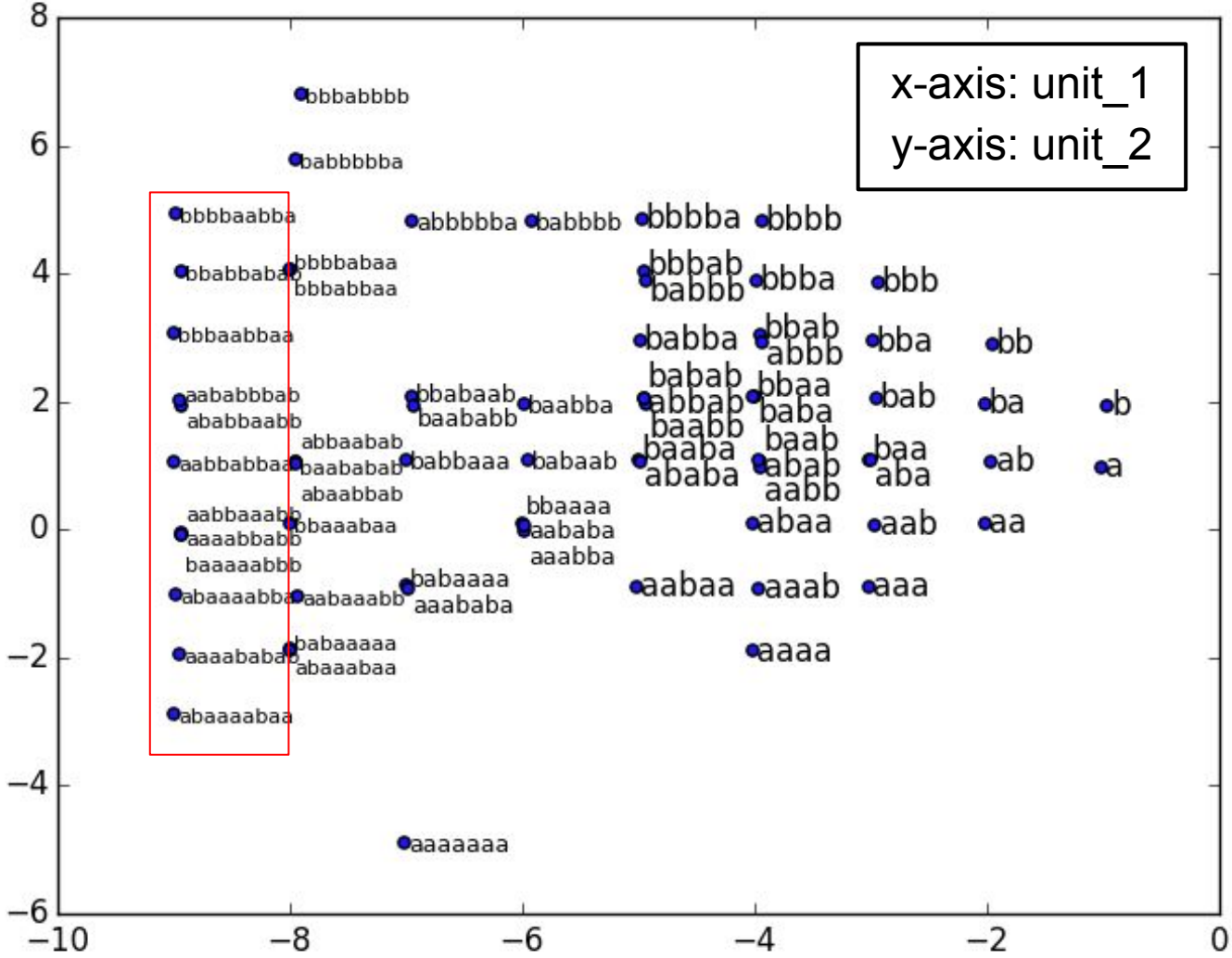


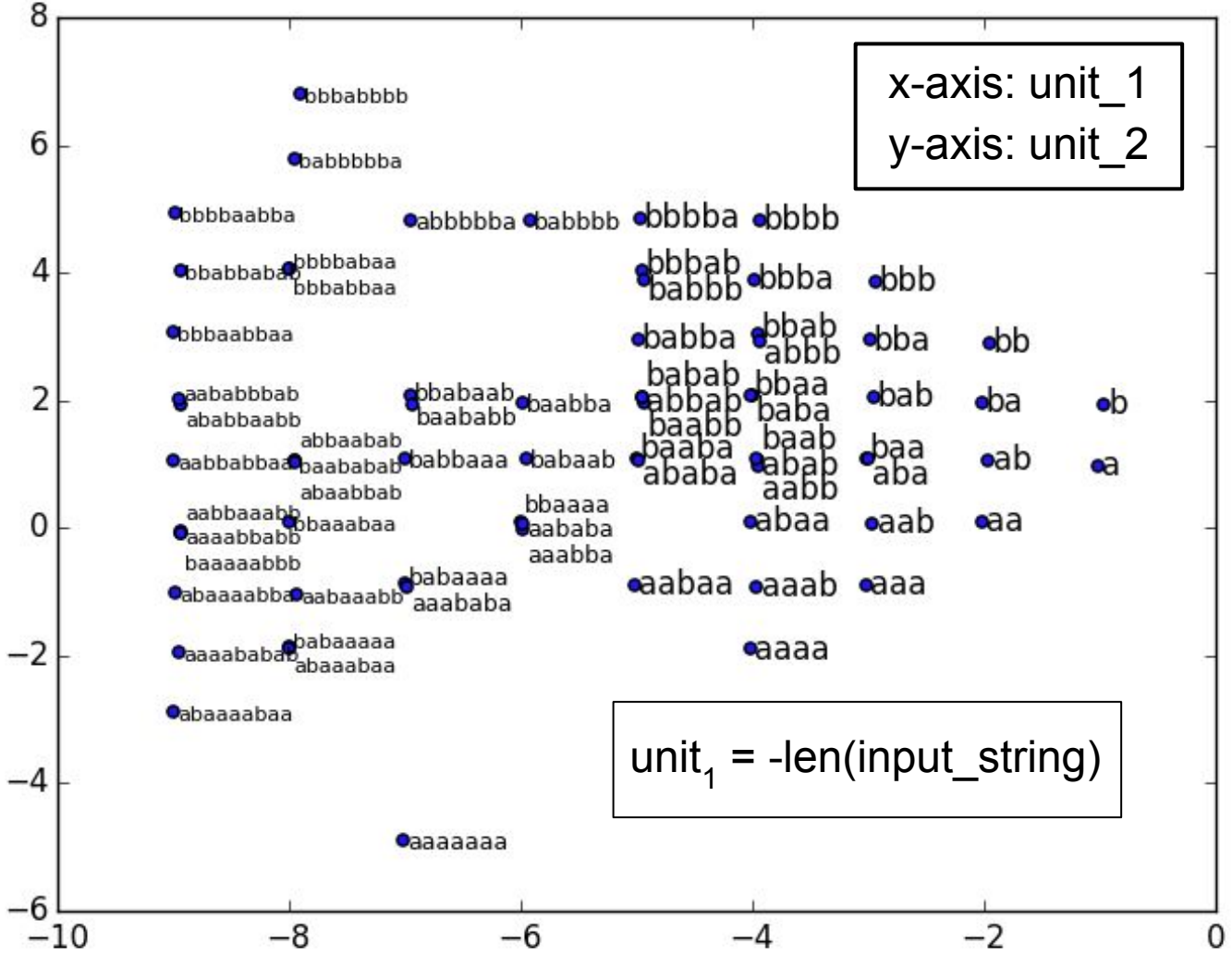
C_t involves only **elementwise** + and x.


x-axis: unit_1
y-axis: unit_2
8
6
4
2
0
−2
−4
−6
−10
−8
−6
−4
−2
0
bbbabbbb
babbbbba
bbbbaabba
abbbbba
babbbb
bbbba
bbbb
bbbbabaa
bbbabbaa
bbbab
babbb
bbba
bbb
bbbaabbaa
babba
bbab
abbb
bba
bb
babab
bbaa
aababbbab
bbabaab
baabba
abbab
bab
ba
b
ababbaabb
baababb
baabb
baba
abbaabab
baaba
baab
aabbabbaa
baababab
babbaaa
babaab
ababa
abab
baa
ab
a
abaabbab
aabbaaabb
bbaaaaa
abaa
aab
aa
bbaaabaa
aababa
aaaabbabb
aaabba
baaaaabbb
babaaaa
abaaaabba
aabaaabb
aaababa
aabaa
aaab
aaa
babaaaaa
aaaababab
abaaabaa
aaaa
abaaaabaa
aaaaaaa


x-axis: unit_1
y-axis: unit_2
8
6
4
2
0
−2
−4
−6
−10
−8
−6
−4
−2
0
bbbabbbb
babbbbba
bbbbaabba
abbbbbba
babbbb
bbbba
bbbb
bbbbabaa
bbbabbaa
bbbab
babbb
bbba
bbb
bbbaabbaa
babba
bbab
abbb
bba
bb
babab
bbaa
aababbbab
bbabaab
baabba
abbab
bab
ba
b
ababbaabb
baababb
baabb
baba
abbaabab
baaba
baab
aabbabbaa
baababab
babbaaa
babaab
ababa
abab
baa
ab
a
abaabbab
aabbaaabb
bbaaaaa
abaa
aab
aa
bbaaabaa
aababa
aaaabbabb
aaabba
baaaaabbb
babaaaa
abaaaabba
aabaaabb
aabaa
aaab
aaa
aaababa
babaaaaa
aaaa
aaaababab
abaaabaa
abaaaabaa
aaaaaaa

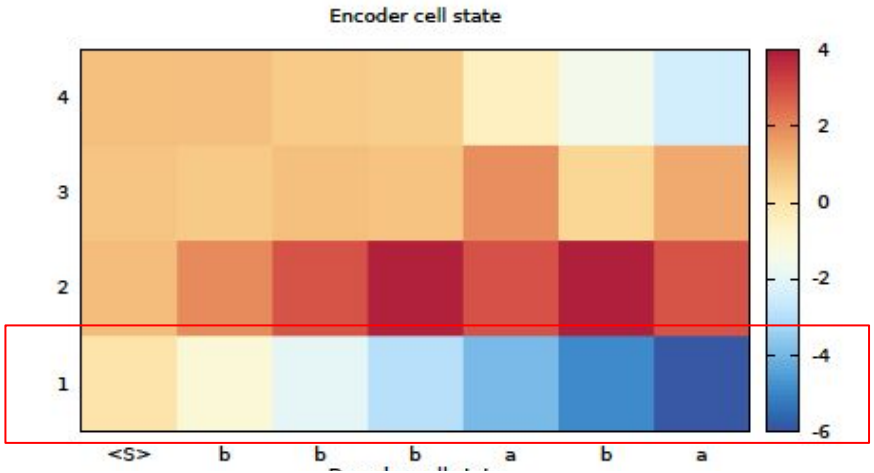

x-axis: unit_1
y-axis: unit_2
8
6
4
2
0
−2
−4
−6
−10
−8
−6
−4
−2
0
bbbabbbb
babbbbba
bbbbaabba
abbbbba
babbbb
bbbba
bbbb
bbbbabaa
bbbabbaa
bbbab
babbb
bbba
bbb
bbbaabbaa
babba
bbab
abbb
bba
bb
babab
aababbbab
bbabaab
baabba
abbab
bbaa
bab
ba
b
ababbaabb
baababb
baabb
baba
abbaabab
baaba
baab
aabbabbaa
baababab
babbaaa
babaab
ababa
abab
baa
ab
a
abaabbab
aba
aabbaaabb
bbaaaa
abaa
aab
aa
bbaaabaa
aababa
aaaabbabb
aaabba
baaaaabbb
babaaaa
abaaaabba
aabaaabb
aabaa
aaab
aaa
aaababa
babaaaaa
aaaababab
aaaa
abaaaabaa
abaaabaa
aaaaaaa


x-axis: unit_1
y-axis: unit_2
8
6
4
2
0
−2
−4
−6
−10
−8
−6
−4
−2
0
bbbabbbb
babbbbba
bbbbaabba
abbbbba
babbbb
bbbba
bbbb
bbbbabaa
bbbabbaa
bbbab
babbb
bbba
bbb
bbbaabbaa
babba
bbab
abbb
bba
bb
aababbbab
ababbaabb
bbabaab
baababb
baabba
babab
abbab
bbaa
baabb
baba
bab
ba
b
abbaabab
aabbabbaa
baababab
babbaaa
babaab
baaba
baab
ababa
abab
baa
aba
ab
a
abaabbab
aabbaaabb
aaaabbabb
bbaaaaa
bbaaaba
aababa
aaabba
abaa
aab
aa
baaaaabbb
abaaaabba
aabaaabb
babaaaa
aaababa
aabaa
aaab
aaa
aaaa
aaaababa
babaaaaa
abaaabaa
abaaaabaa
aaaaaaa


x-axis: unit_1
y-axis: unit_2
unit1 = -len(input_string)
bbbabbbb
babbbbba
bbbbaabba
abbbbba
babbbb
bbbba
bbbb
bbbbabaa
bbbabbaa
bbbab
babbb
bbba
bbb
bbbaabbaa
babba
bbab
abbb
bba
bb
babab
bbaa
aababbbab
bbabaab
baabba
abbab
bab
ba
b
ababbaabb
baababb
baabb
baab
abbaabab
baaba
baa
aabbabbaa
baababab
babbaaa
babaab
ababa
abab
ab
a
abaabbab
aba
aabbaaabb
bbaaaaa
abaa
aab
aa
bbaaabaa
aababa
aaaabbabb
aaabba
baaaaabbb
babaaaa
abaaaabba
aabaaabb
aaababa
aabaa
aaab
aaa
babaaaaa
aaaababab
abaaabaa
aaaa
abaaaabaa
aaaaaaa
8
6
4
2
0
−2
−4
−6
−10
−8
−6
−4
−2
0

Toy Example: String Copy

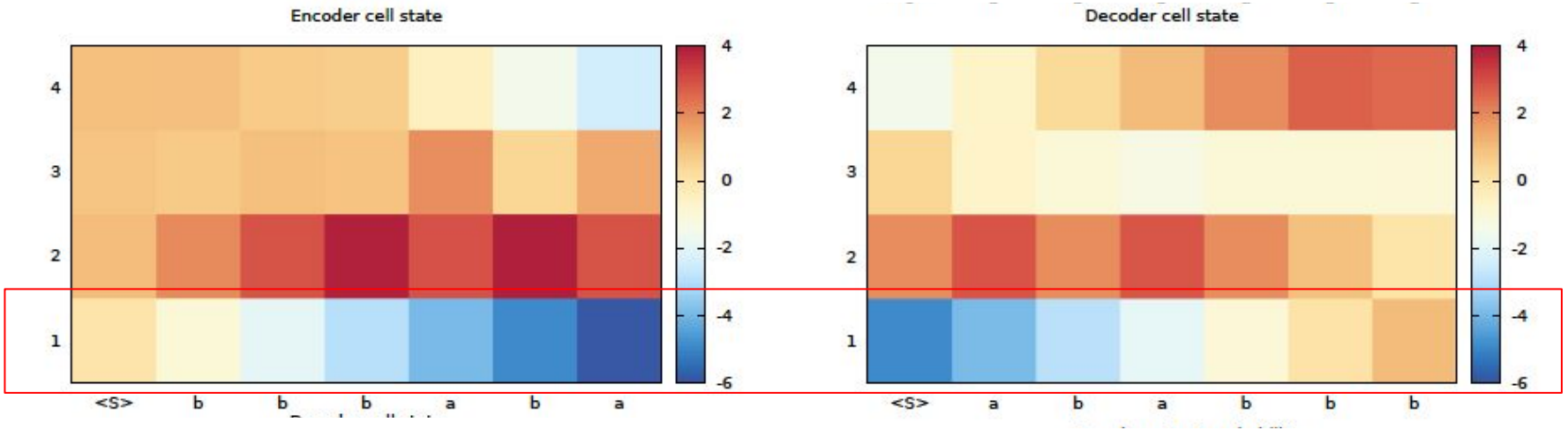
<s> *b b b a b a* → <s> *b b b a b a <EOF>*



Encoding Cell State
unit_1 decrease by 1.0

Toy Example: String Copy

<s> b b b a b a → <s> b b b a b a <EOF>



Encoding Cell State
unit_1 decrease by 1.0

Decoding Cell State
unit_1 increase by 1.0

Full Scale NMT

English => French

1000 hidden units LSTM

2 layers

Non-attention

BLEU = 29.8

Full Scale NMT

$$Y = w_1 * X_1 + w_2 * X_2 + \ldots + w_{1000} * X_{1000} + b$$

Sentence_i	It	is	raining	right	now
Y	1	2	3	4	5
X	1000 cell states	1000 cell states	1000 cell states	1000 cell states	1000 cell states

In total 143,379 (Y, X)

Full Scale NMT

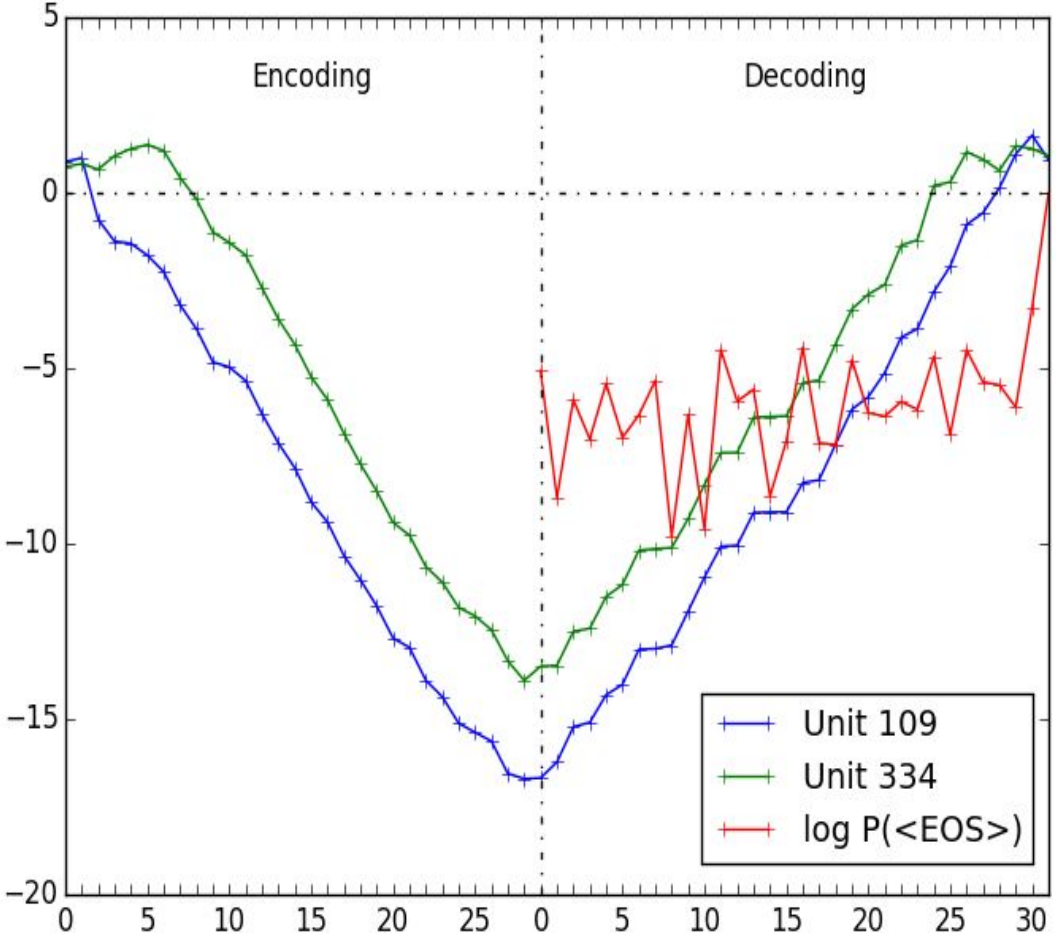
$$Y = w_1 * X_1 + w_2 * X_2 + \ldots + w_{1000} * X_{1000} + b$$

	R^2
1000 units in lower-layer	0.990
1000 units in upper-layer	0.981

Full Scale NMT

k	Best subset of LSTM's 1000 units	R^2
1	109	0.894
2	334, 109	0.936
3	334, 442, 109	0.942
4	334, 442, 109, 53	0.947
5	334, 442, 109, 53, 46	0.951
6	334, 442, 109, 53, 46, 928	0.953
7	334, 442, 109, 53, 46, 433, 663	0.955

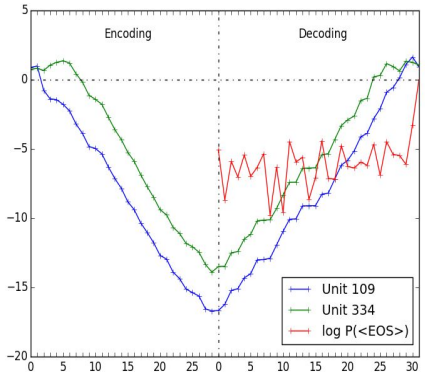
Table 2: Sets of k units chosen by beam search to optimally track length in the NMT encoder. These units are from the LSTM's second layer.



Encoding
Unit 109 and 334 decrease from above zero

Decoding
Increase during decoding, once they are above zero, the model is ready to generate <EOS>.

Conclusion

	Toy Example	**Full Scale NMT**
Who	Unit_1 controls the length	Unit_{109} and Unit_{334} contributes to the length
How		

Thanks and QA