Neural Lattice Search for Domain Adaptation in Machine Translation

Huda Khayrallah, Gaurav Kumar
Kevin Duh, Matt Post, Philipp Koehn

This talk was presented at JHU CLSP seminar
November 17, 2017
It is based on this paper:
http://aclweb.org/anthology/I17-2004
bib:
http://aclweb.org/anthology/I17-2004.bib

Neural Lattice Search for Domain Adaptation in Machine Translation

Huda Khayrallah, Gaurav Kumar
Kevin Duh, Matt Post, Philipp Koehn

Outline

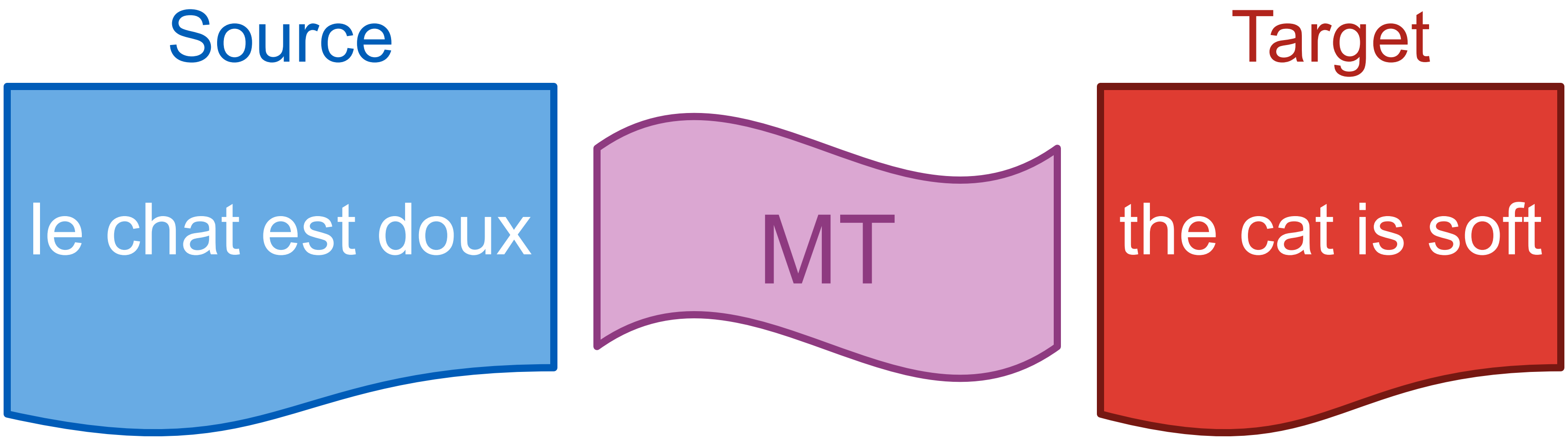
- Machine Translation
  - Phrase Based MT
  - Neural MT
  - MT metrics
  - Domain Adaptation
- Neural Lattice Search for Domain Adaptation in Machine Translation

Machine Translation
(in ~20 min)

Huda Khayrallah

17 November 2017

Machine Translation



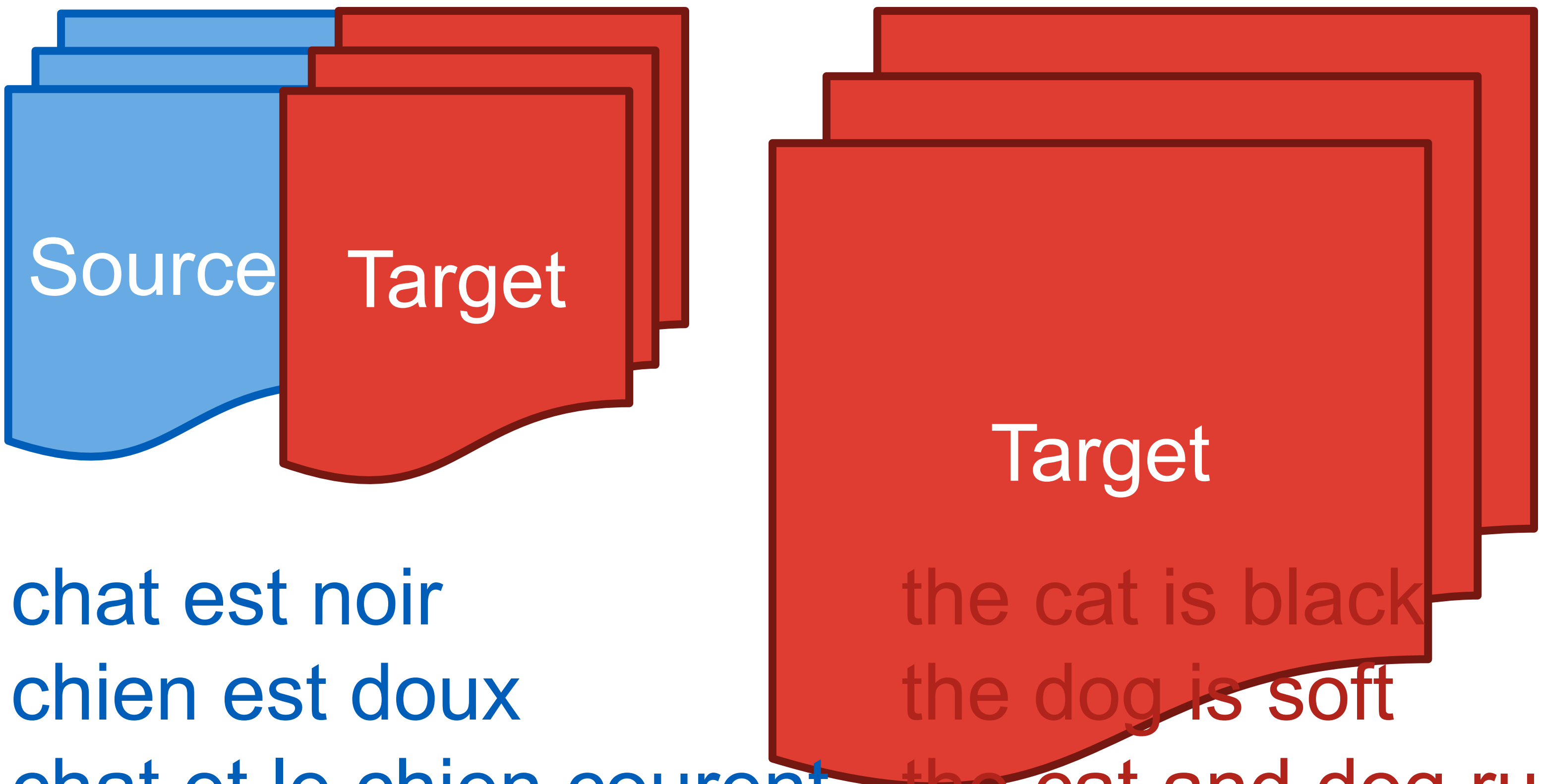
Goals:

Adequacy & Fluency

What kind of data do we have?

Parallel Text

Monolingual Text



le chat est noir
le chien est doux
le chat et le chien courent

the cat is black
the dog is soft
the cat and dog run

What kind of systems do we build?

- Phrase-based MT
- Neural MT

Phrase-Based MT

- Source is segmented in to ‘phrases’
- Phrases translated into target language
- Phrases are reordered

[Koehn et al. 2003]

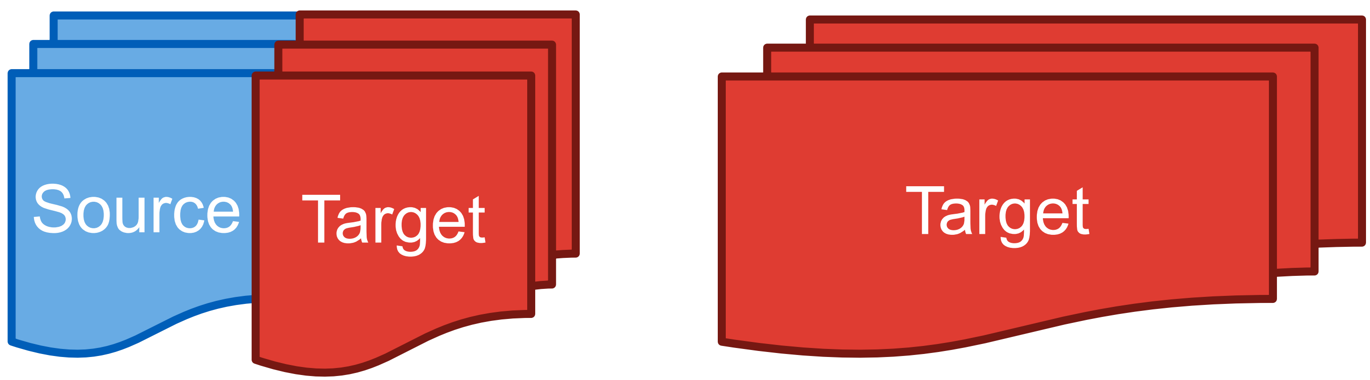
Phrase-Based MT

$$P(target|source)$$

$$\sim \boxed{P(source|target)}\,\boxed{P(target)}$$

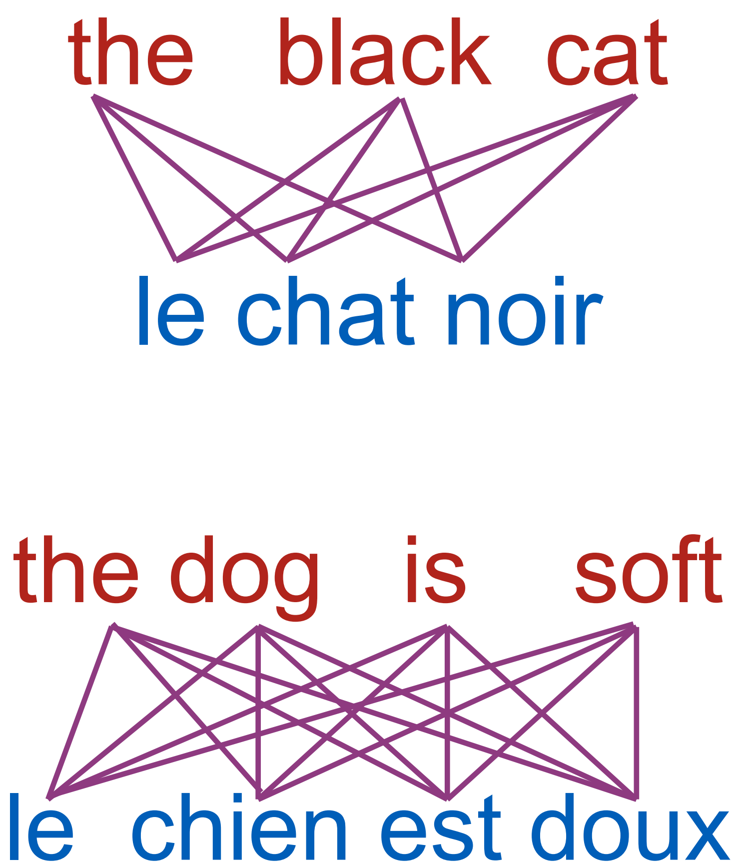
Translation model

Language model

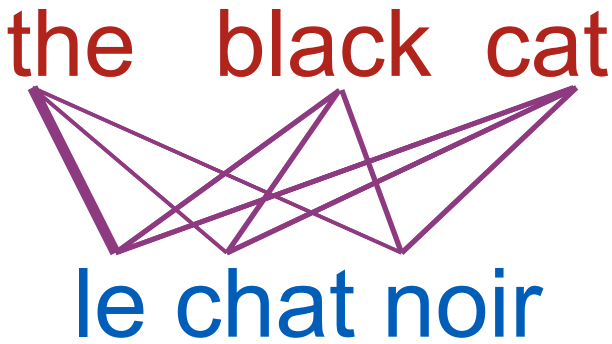


Translation Model

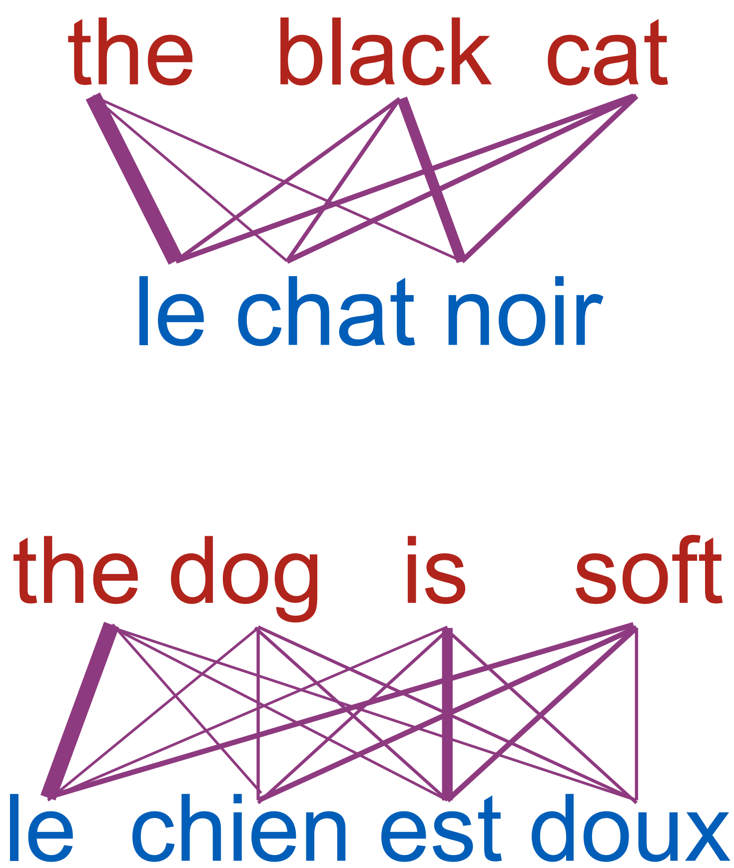
Word Alignment



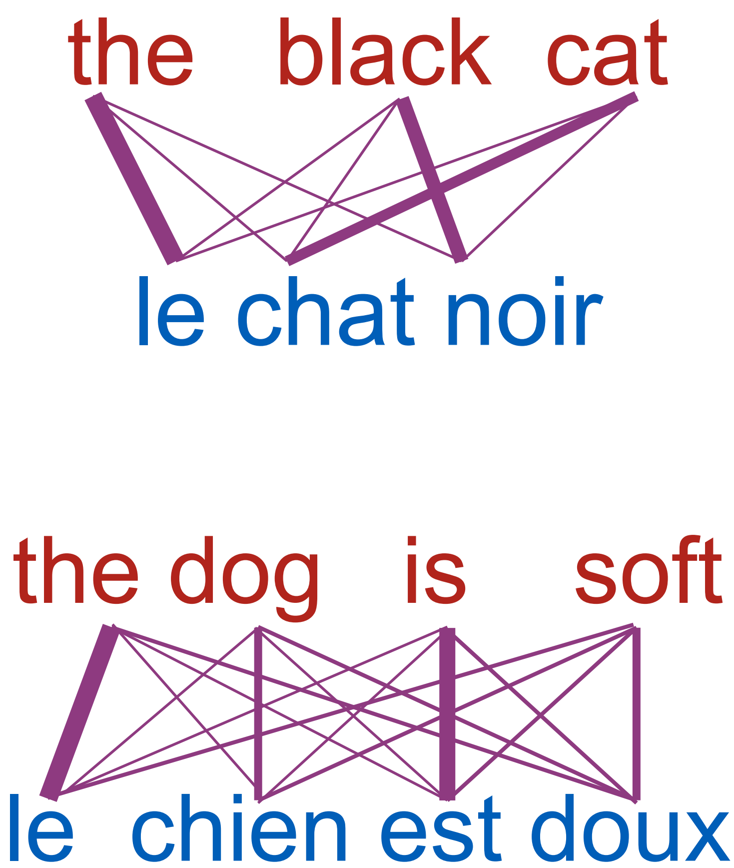
Word Alignment



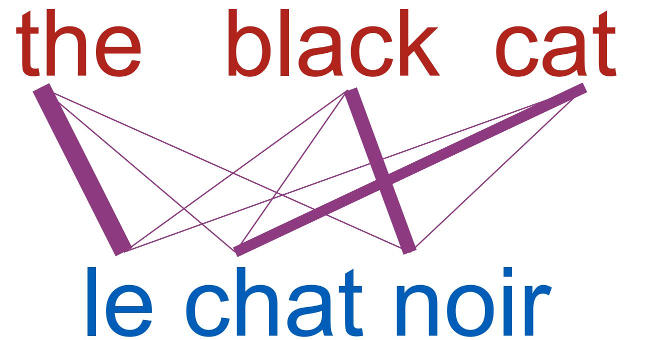
Word Alignment

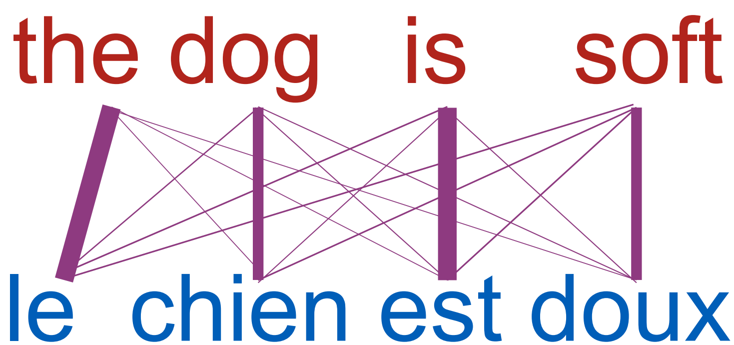


Word Alignment

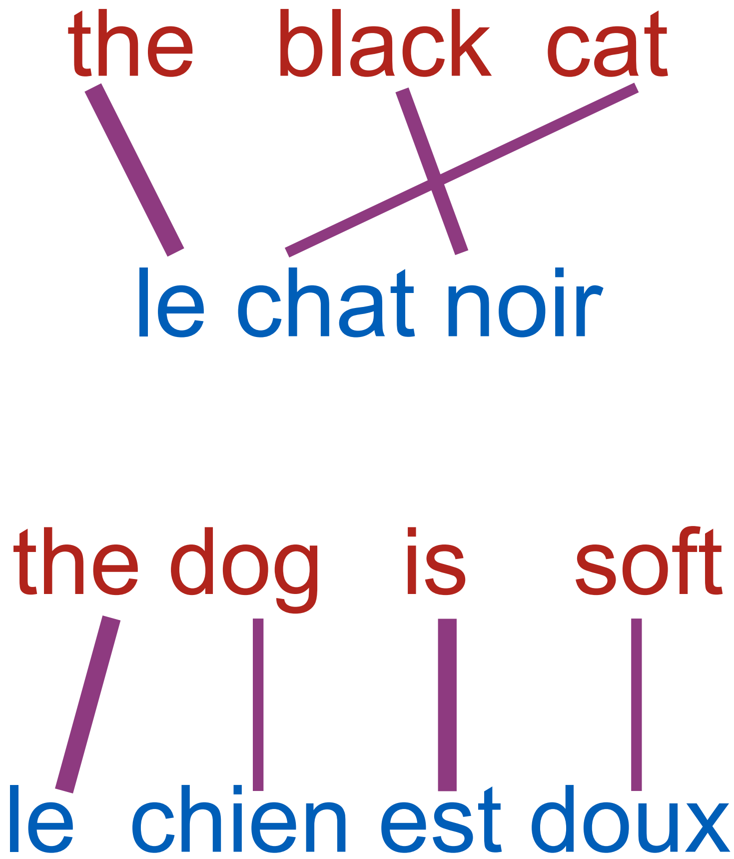


Word Alignment

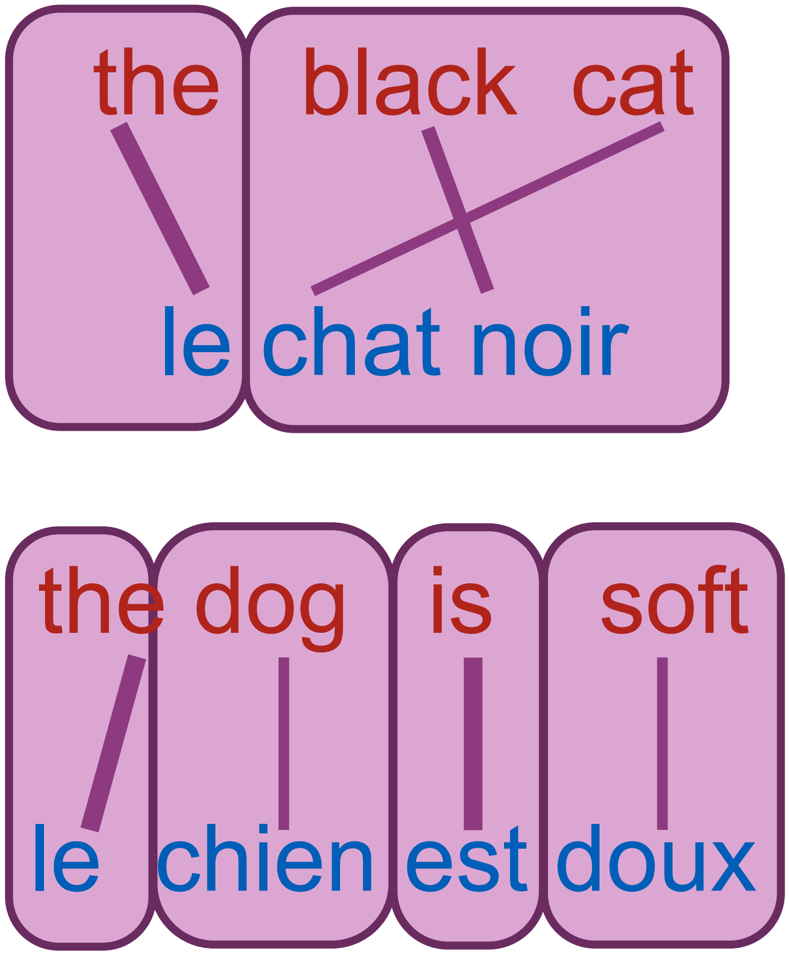




Word Alignment



Phrase Extraction



Language Model

$$P(the\ cat\ is\ soft) > P(the\ cat\ are\ soft)$$
$$P(the\ cat\ is\ soft) > P(the\ cat\ soft\ is)$$

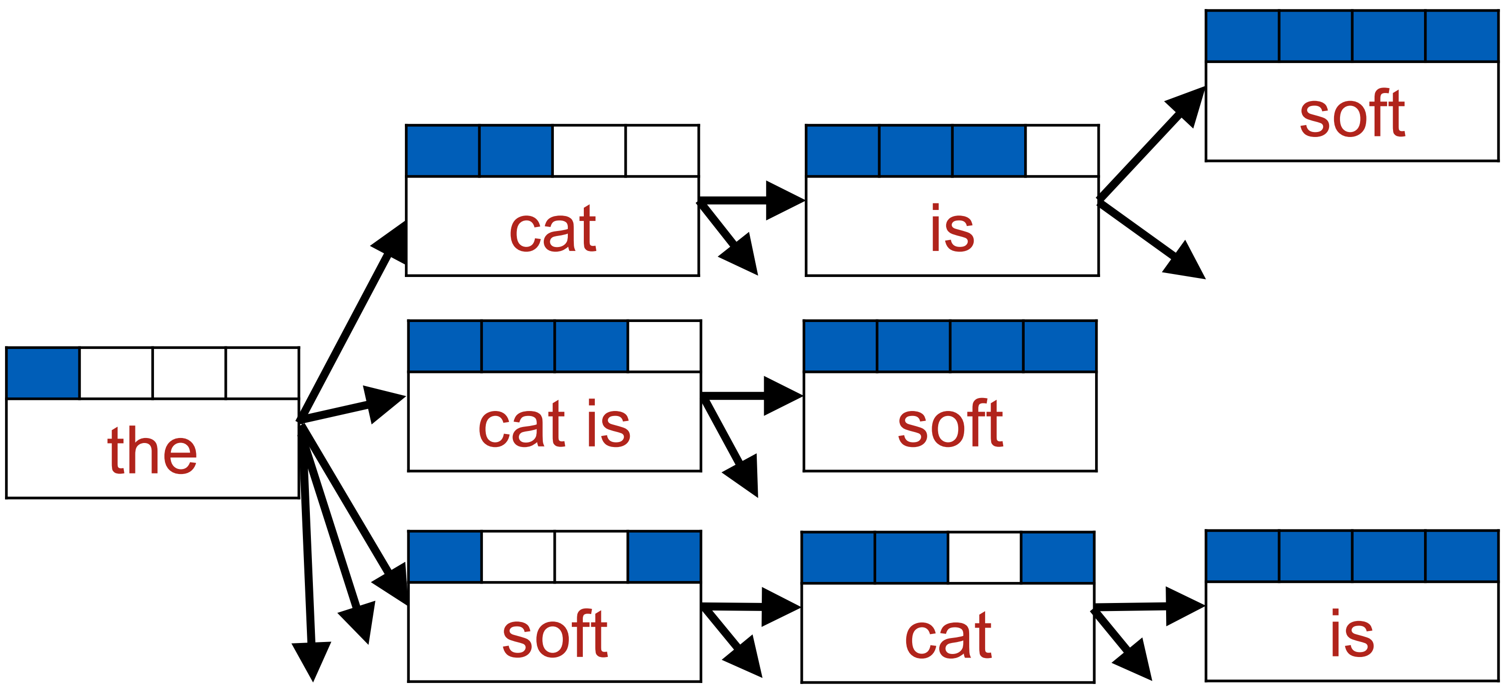
- *n*-gram $(\mathrm{n} = 5)$

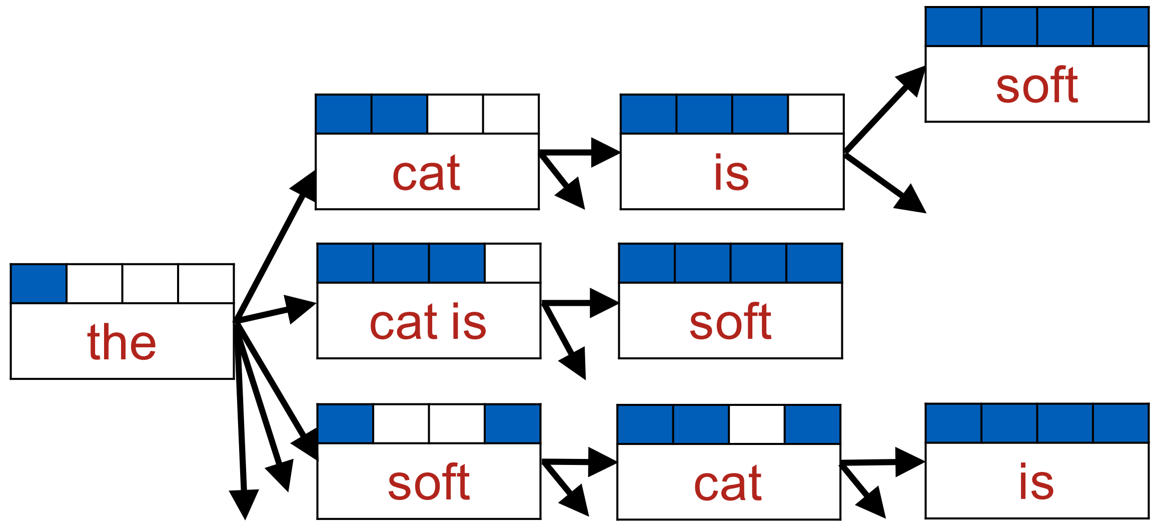
Decoding

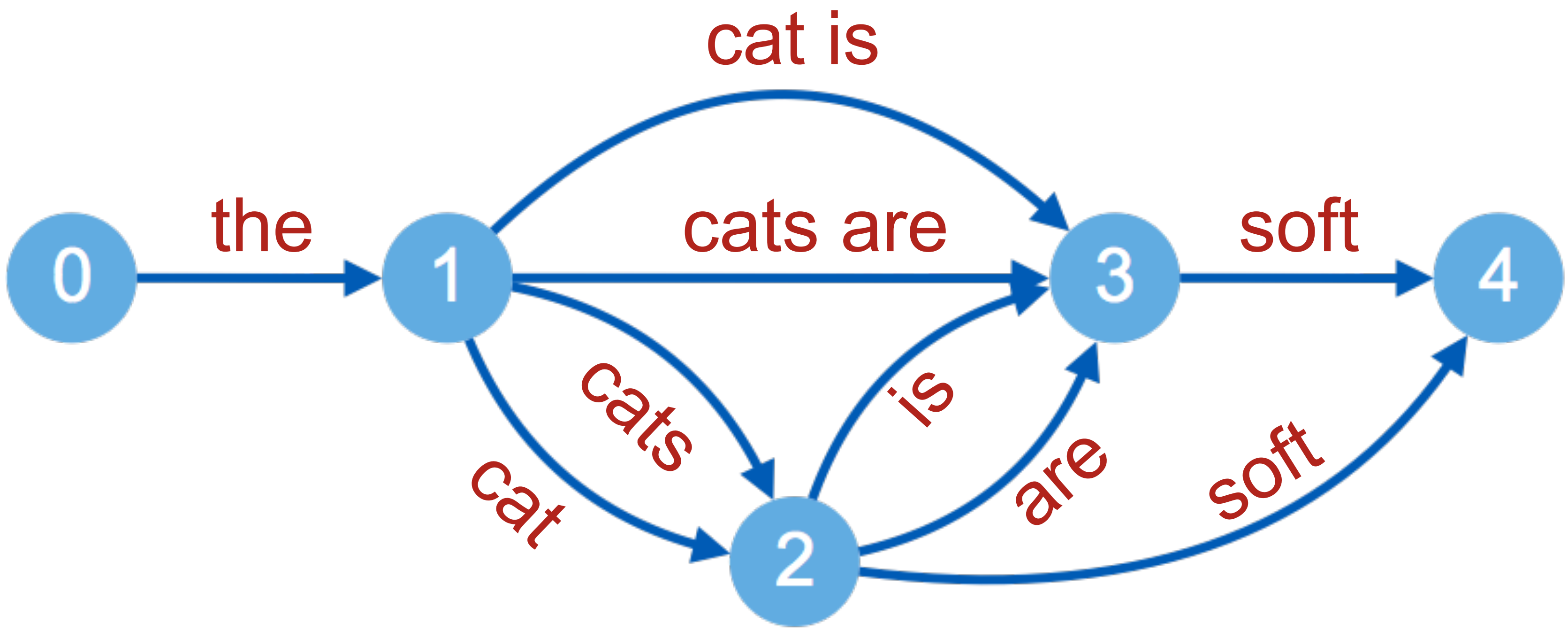
<table>
<tr><td>le</td><td>chat</td><td>est</td><td>doux</td></tr>
<tr><td>the</td><td>cat</td><td>is</td><td>soft</td></tr>
<tr><td>a</td><td colspan="2">cat is</td><td>sweet</td></tr>
<tr><td></td><td>cats</td><td>are</td><td></td></tr>
<tr><td></td><td colspan="2">cats are</td><td></td></tr>
</table>

Decoding

<table>
<tr><th>le</th><th>chat</th><th>est</th><th>doux</th></tr>
<tr><td>the</td><td>cat</td><td>is</td><td>soft</td></tr>
<tr><td>a</td><td colspan="2">cat is</td><td>sweet</td></tr>
<tr><td></td><td>cats</td><td>are</td><td></td></tr>
<tr><td></td><td colspan="2">cats are</td><td></td></tr>
</table>



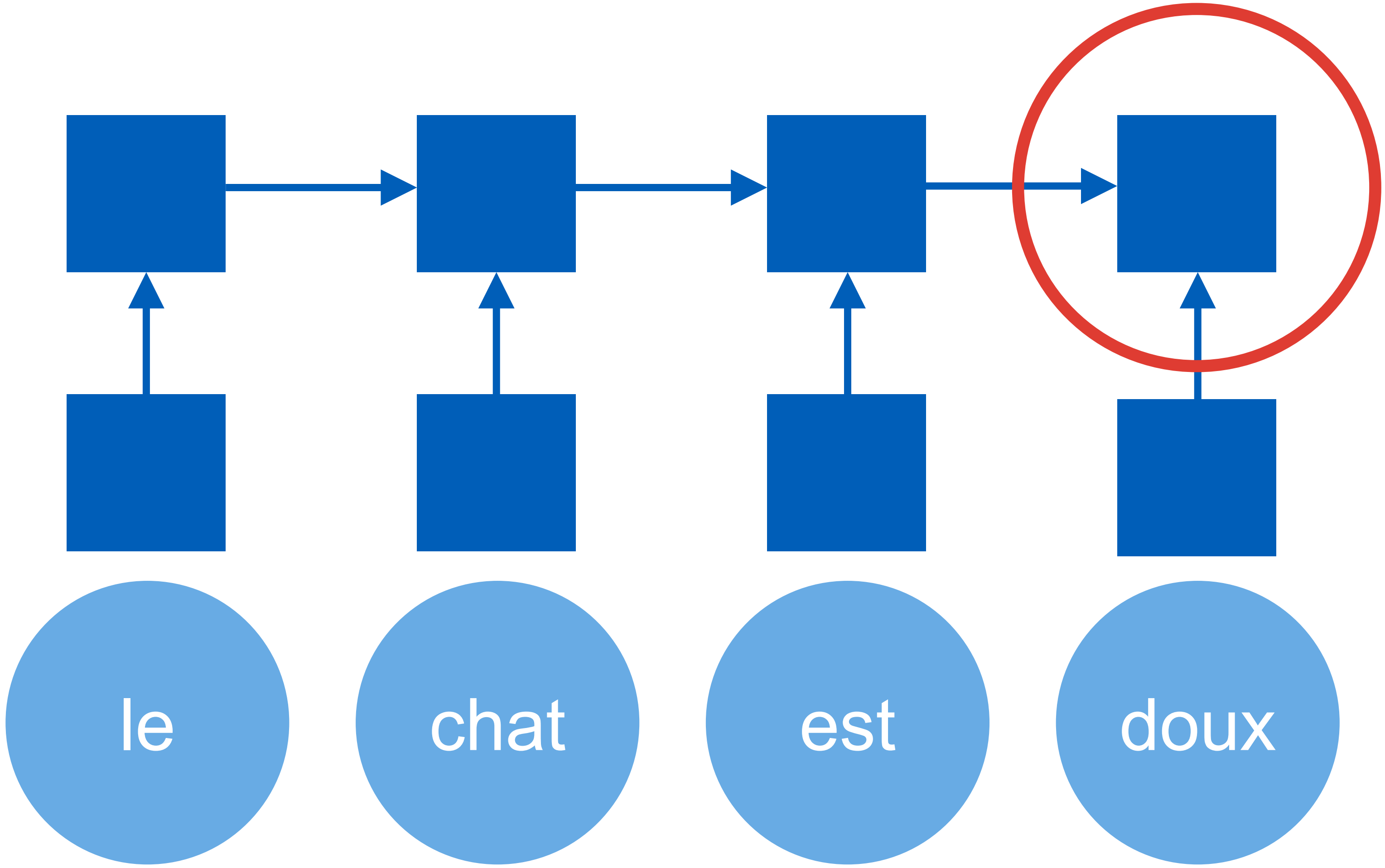

the
cat
cat is
soft
is
soft
soft
cat
is


0
1
2
3
4
the
cat is
cats are
soft
cats
cat
is
are
soft

Neural MT

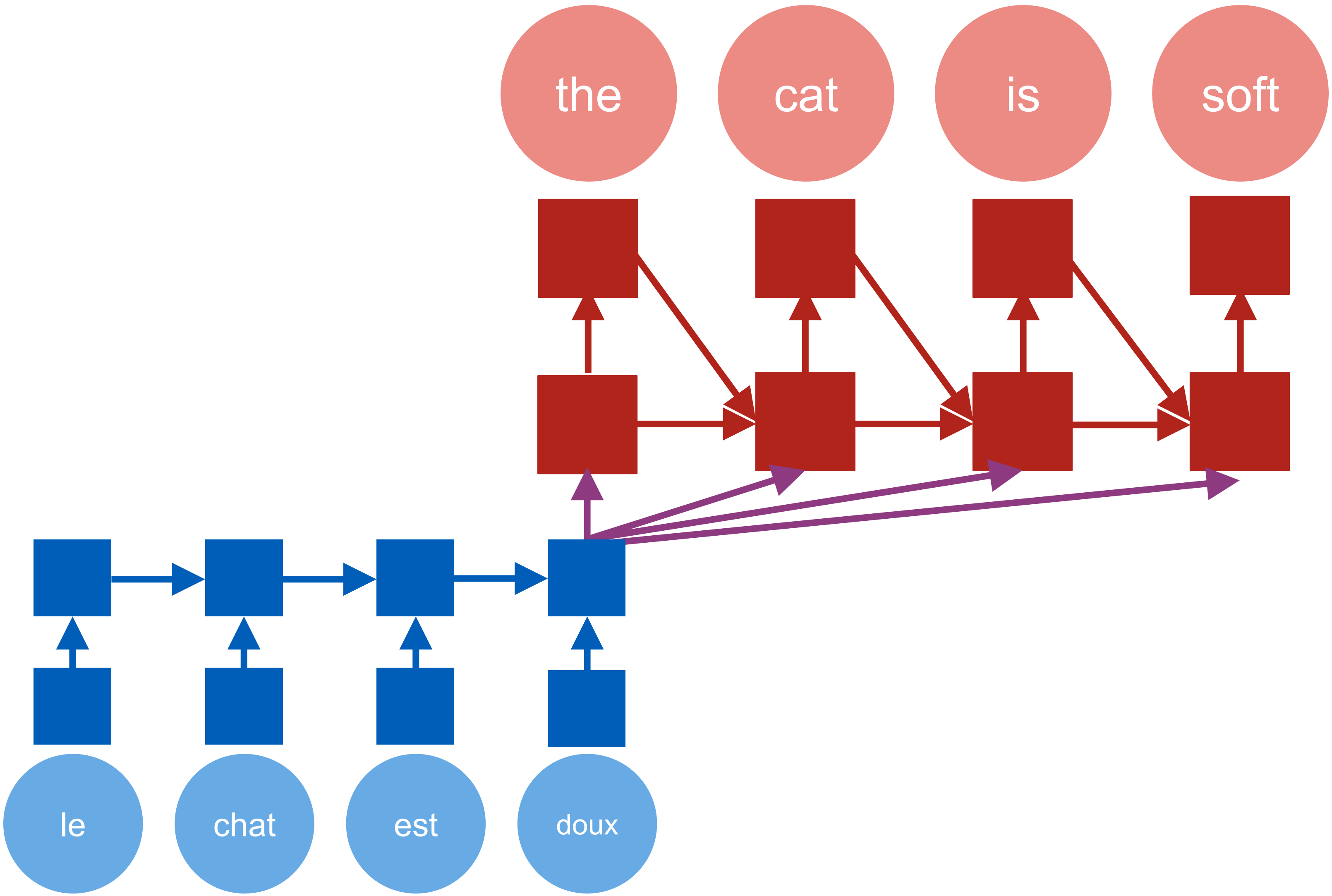
- Each word is represented as a vector
- Recurrent neural net encoder and decoder

Neural MT



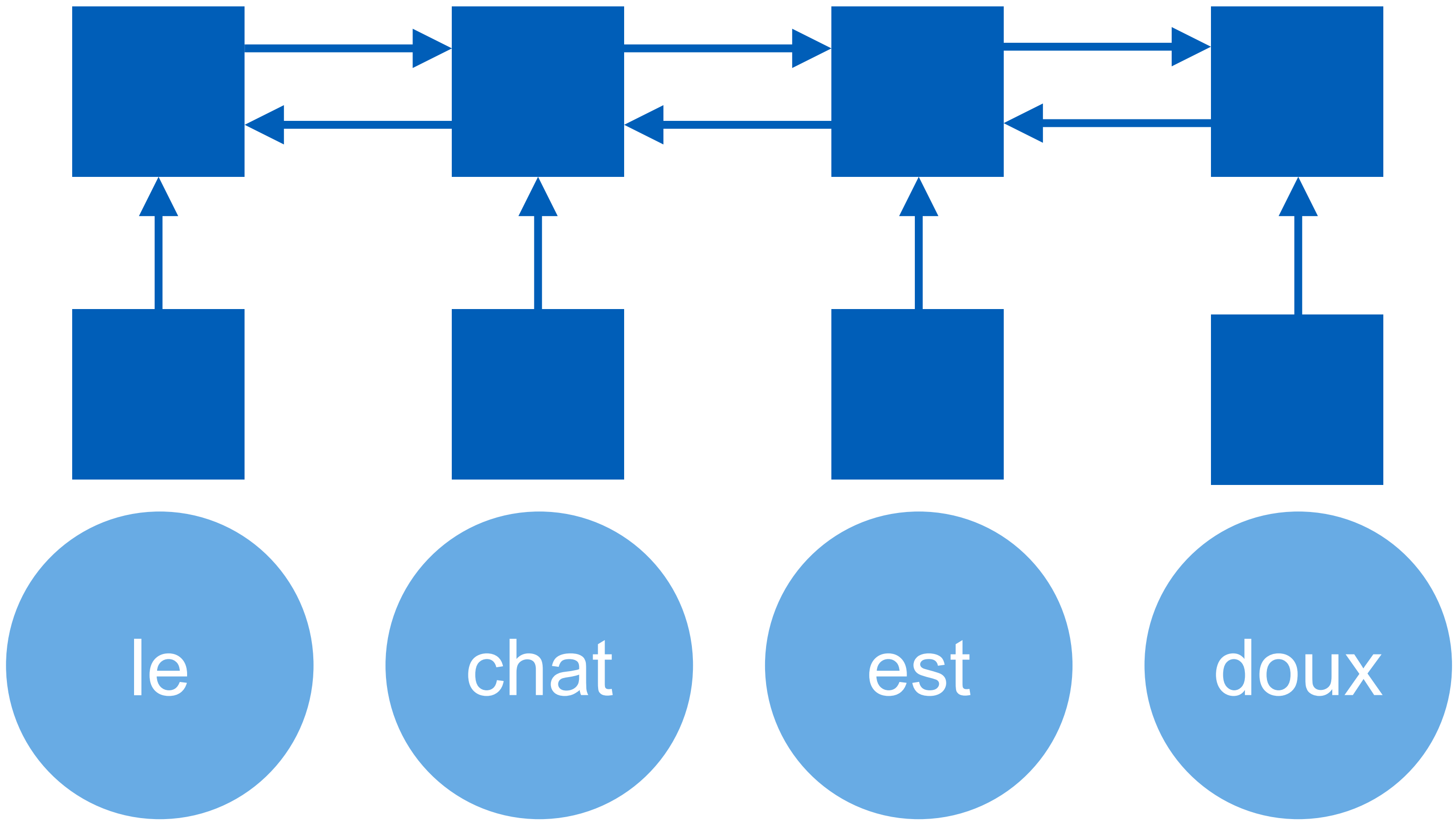
[Sutskever et al. 2015]

Neural MT



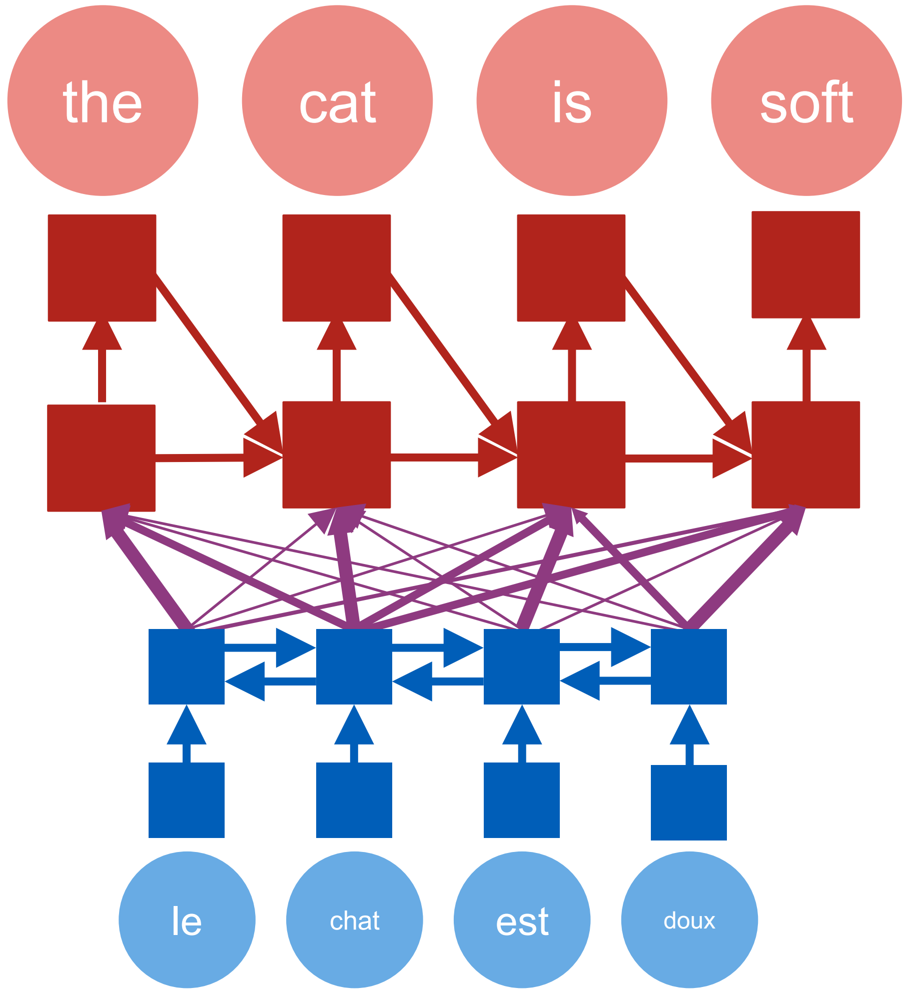
[Sutskever et al. 2015]

(bidirectional) Neural MT



[Schuster & Paliwal 1997]

Neural MT (with attention)



[Bahdanau et al. 2015]

Differences

PBMT

– 5-gram history

– Explicit translation for each phrase

Adequacy?

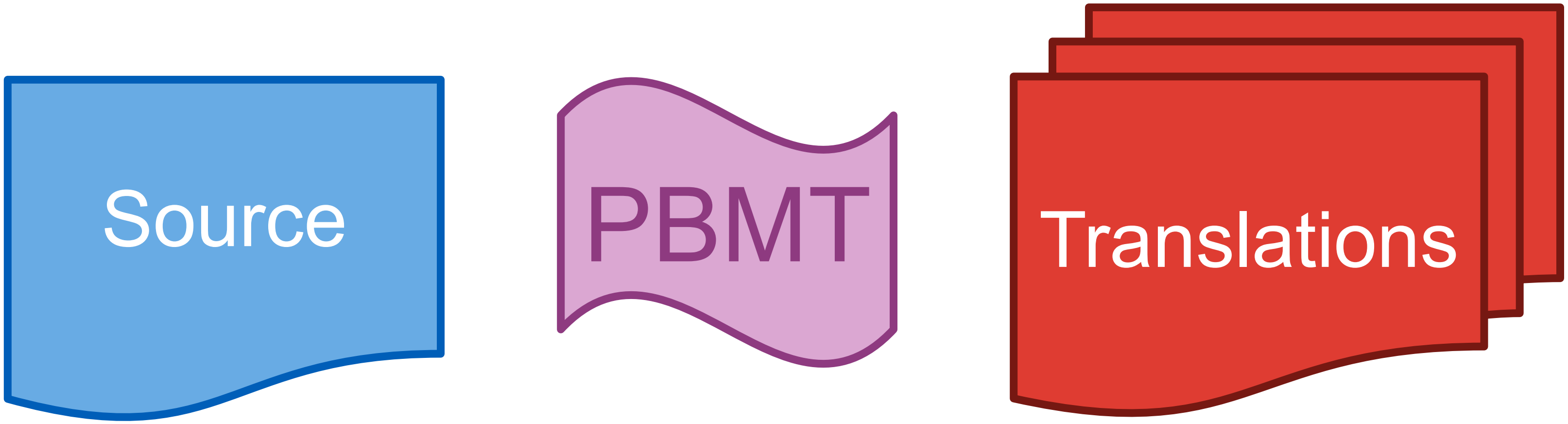
NMT

– Full sentence history

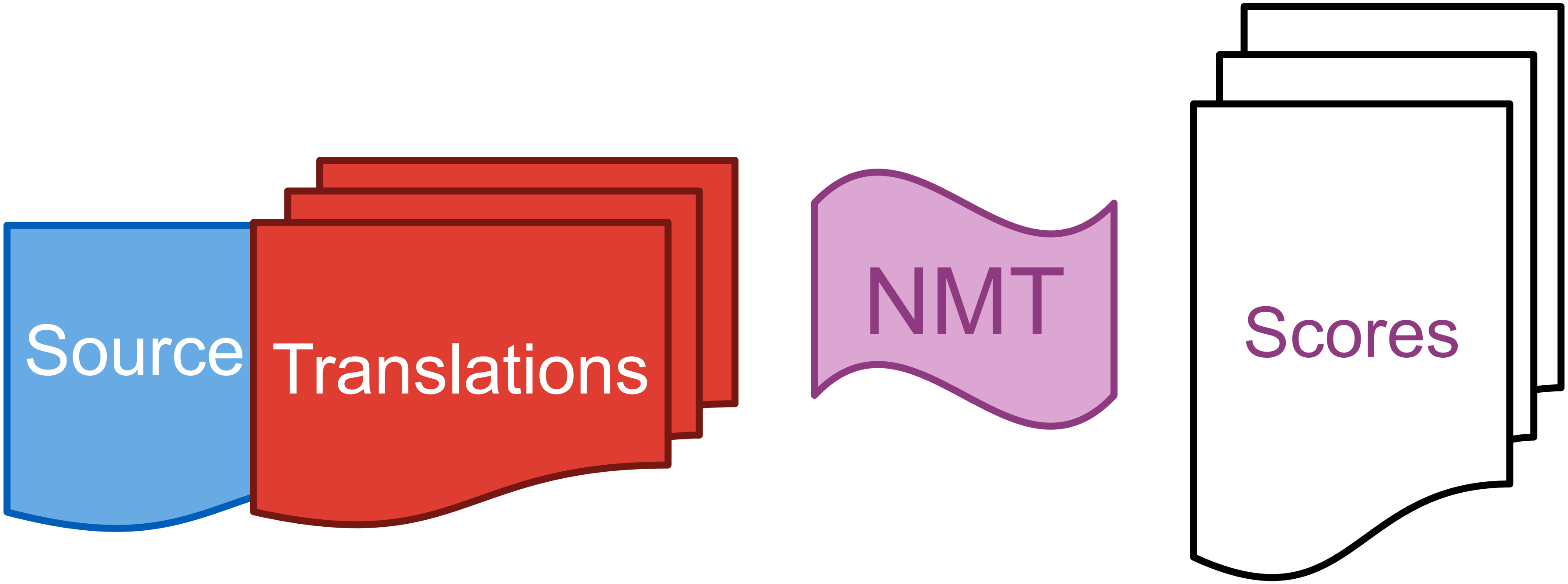
Fluency?

– Soft ‘alignment’

n-best rescoring



n-best rescoring



MT Metrics

- Metric should be:
  - Meaningful
  - Correct
  - Consistent
  - Low Cost
  - Useable for tuning

Why not WER?

- WER ≈ Edit Distance
- Does not account for reordering
- Exact match is not a great goal
  - Should the Ice Bucket Challenge be forbidden?
  - This "ice bucket" could be banned then?
  - Now are they going to ban the ice bucket?

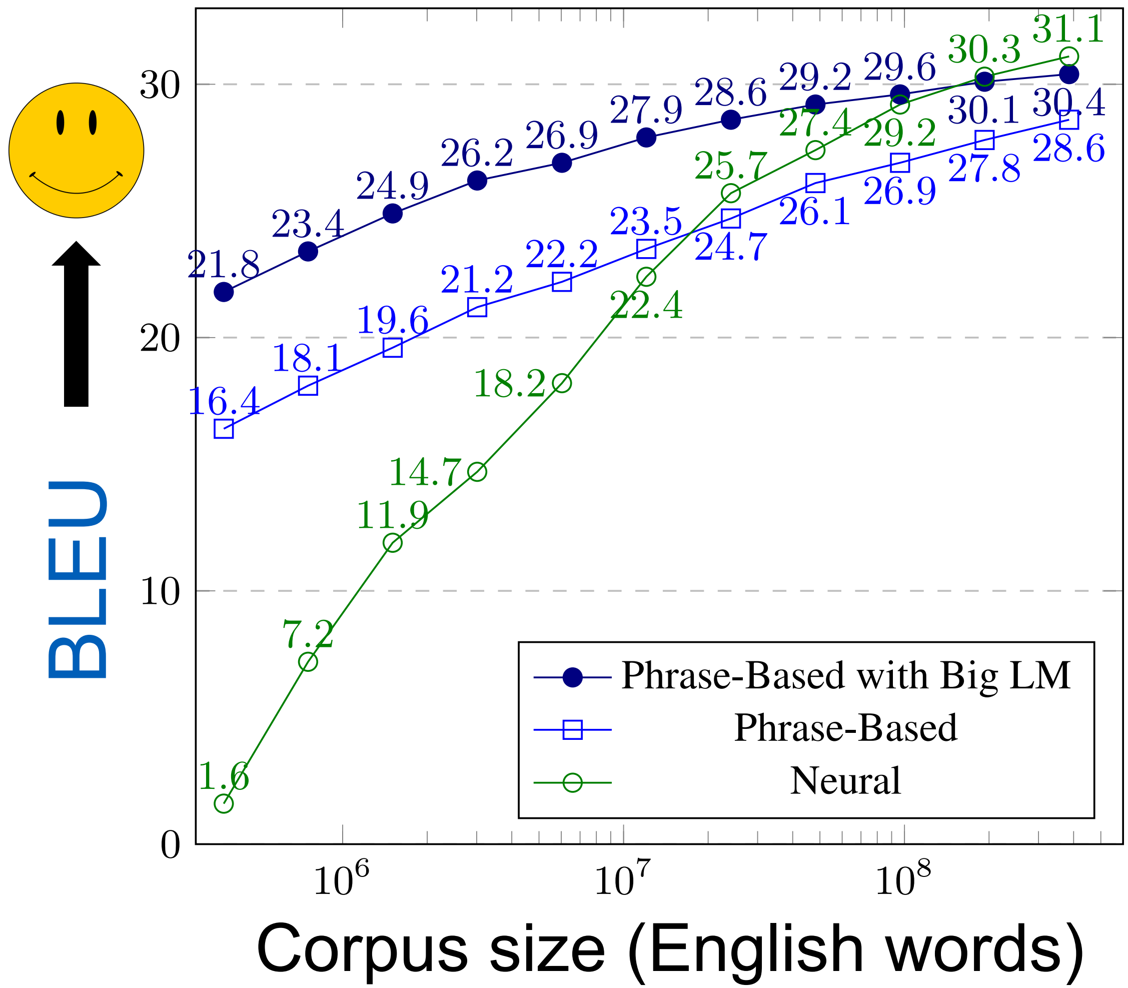
BLEU

- Weighted *n*-grams precision

$$min\left(1, \frac{output\ length}{reference\ length}\right)\left(\prod_{i=1}^{4} precision_i\right)^{\frac{1}{4}}$$

- Between 0 and 1
  - (often scaled to be 0-100)
- Higher is better
- Imperfect…
- But… not bad

How much text do we need?

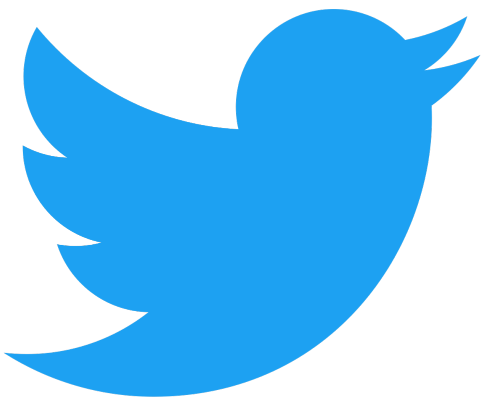


[Koehn & Knowles 2017]

Where does parallel text come from?

Would it not be beneficial, in the short term, following the Rotterdam model, to inspect according to a points system in which, for example, account is taken of the ship's age, whether it is single or double-hulled or whether it sails under a flag of convenience.

What do we want to translate?



Maritza RITZ Willis
@RitzWillis

Follow

I have 2 children with me and tge,water is swallowing us up. Please send help.
911 is not responding!!!!!!

2:07 AM - 27 Aug 2017

15,416 Retweets **8,583** Likes

477 15K 8.6K

Developmental toxicity, including dose-dependent delayed foetal ossification and possible teratogenic effects, were observed in rats at doses resulting in subtherapeutic exposures (based on AUC) and in rabbits at doses resulting in exposures 3 and 11 times the mean steady-state AUC at the maximum recommended clinical dose.

Annual growth in prices came in at 10.9 per cent, more than double the gain of the 12 months to August 2013, but the gains were not evenly spread across the country

Domain adaptation:

- When train and test differ

Domain adaptation (in practice):

- Large amount of out-of-domain data
- Small amount of in-domain data

Neural Lattice Search for Domain Adaptation in Machine Translation

Huda Khayrallah, Gaurav Kumar
Kevin Duh, Matt Post, Philipp Koehn

combine

adequacy of PBMT

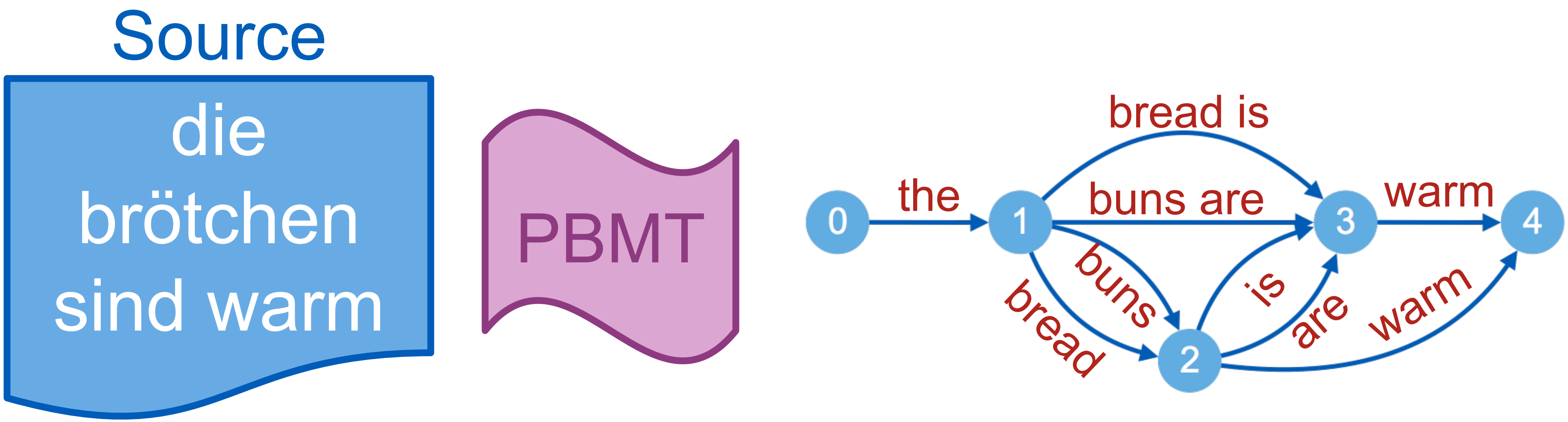
with

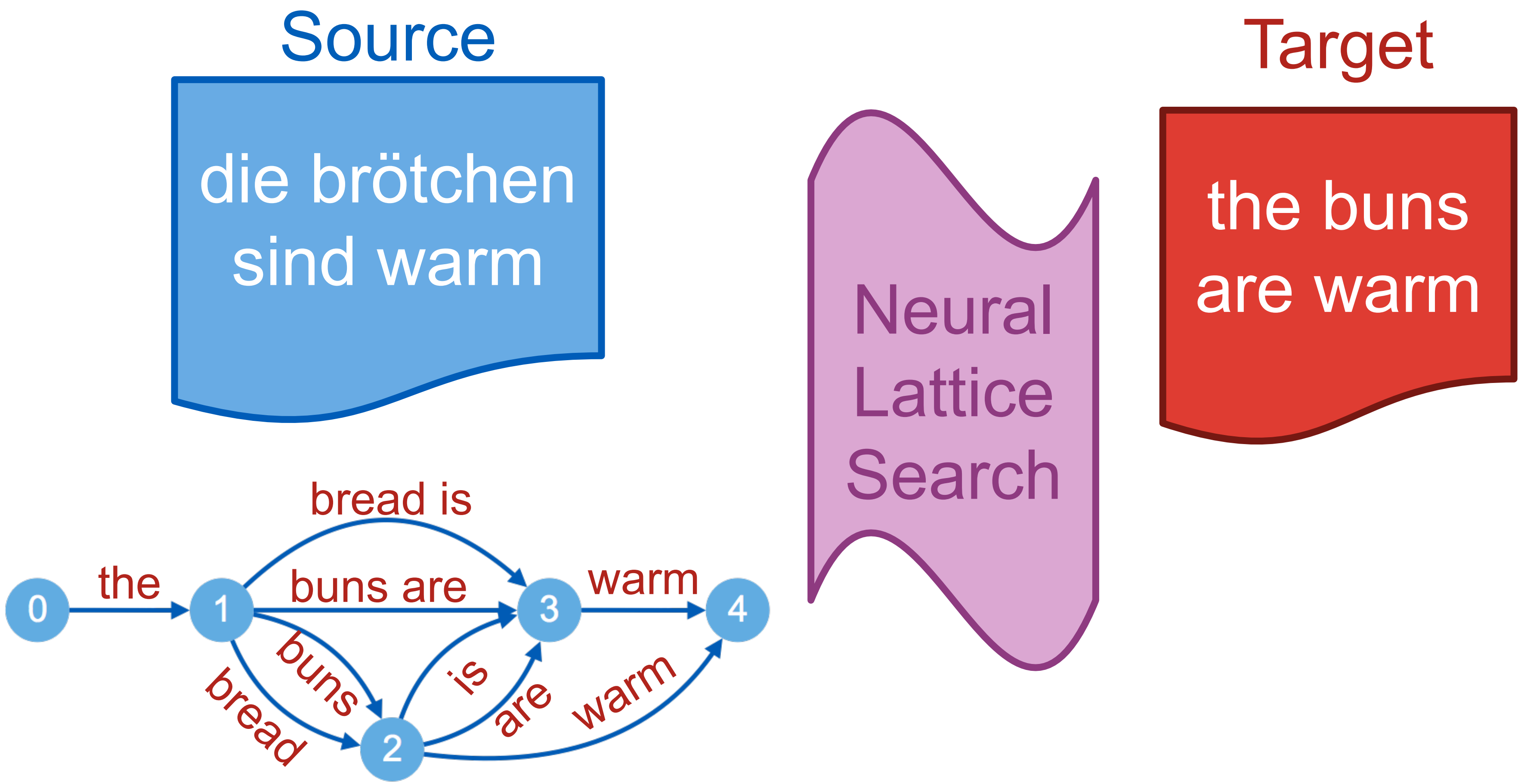
fluency of NMT

use PBMT

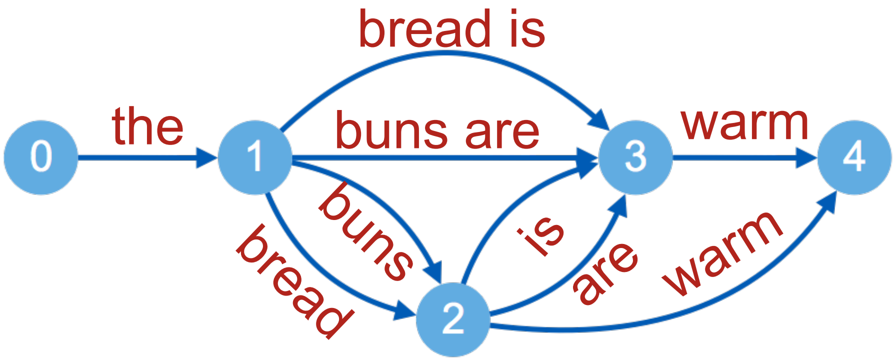
to constrain the search space

of NMT

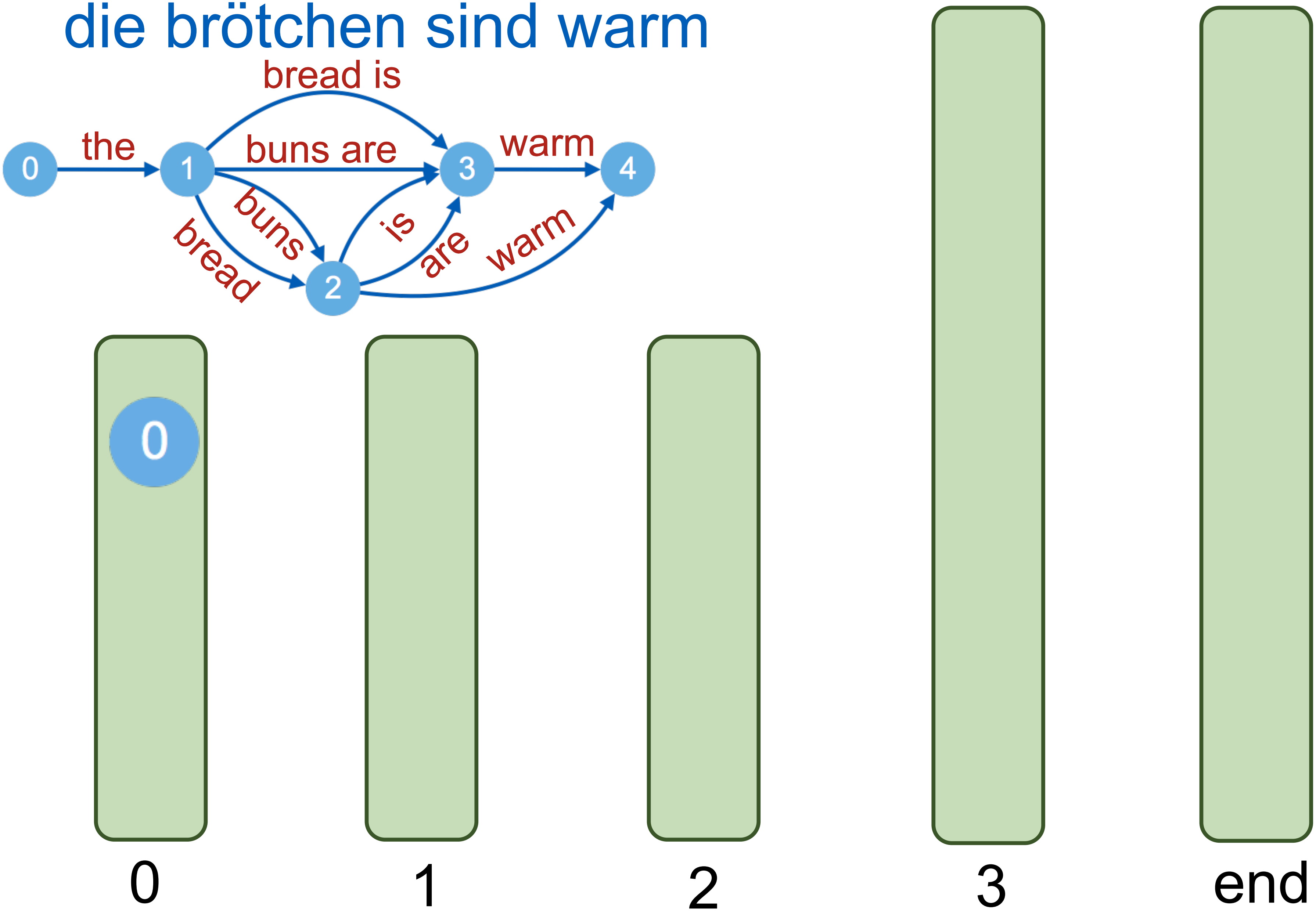

Source
die
brötchen
sind warm
PBMT
0
1
2
3
4
the
bread is
buns are
warm
buns
bread
is
are
warm


Source
die brötchen
sind warm
Neural
Lattice
Search
Target
the buns
are warm
0
the
1
bread is
buns are
3
warm
4
bread
buns
2
is
are
warm

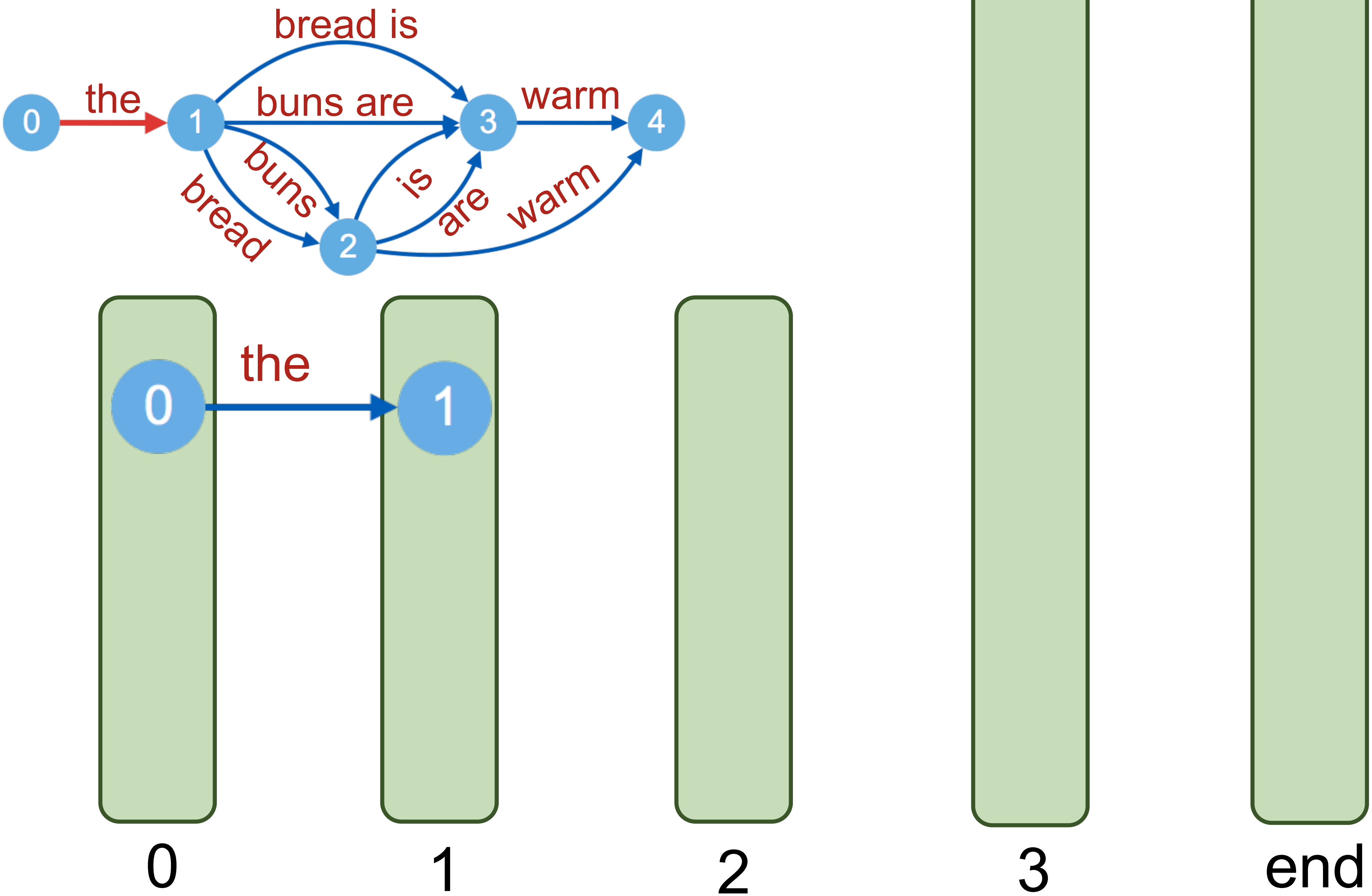
die brötchen sind warm



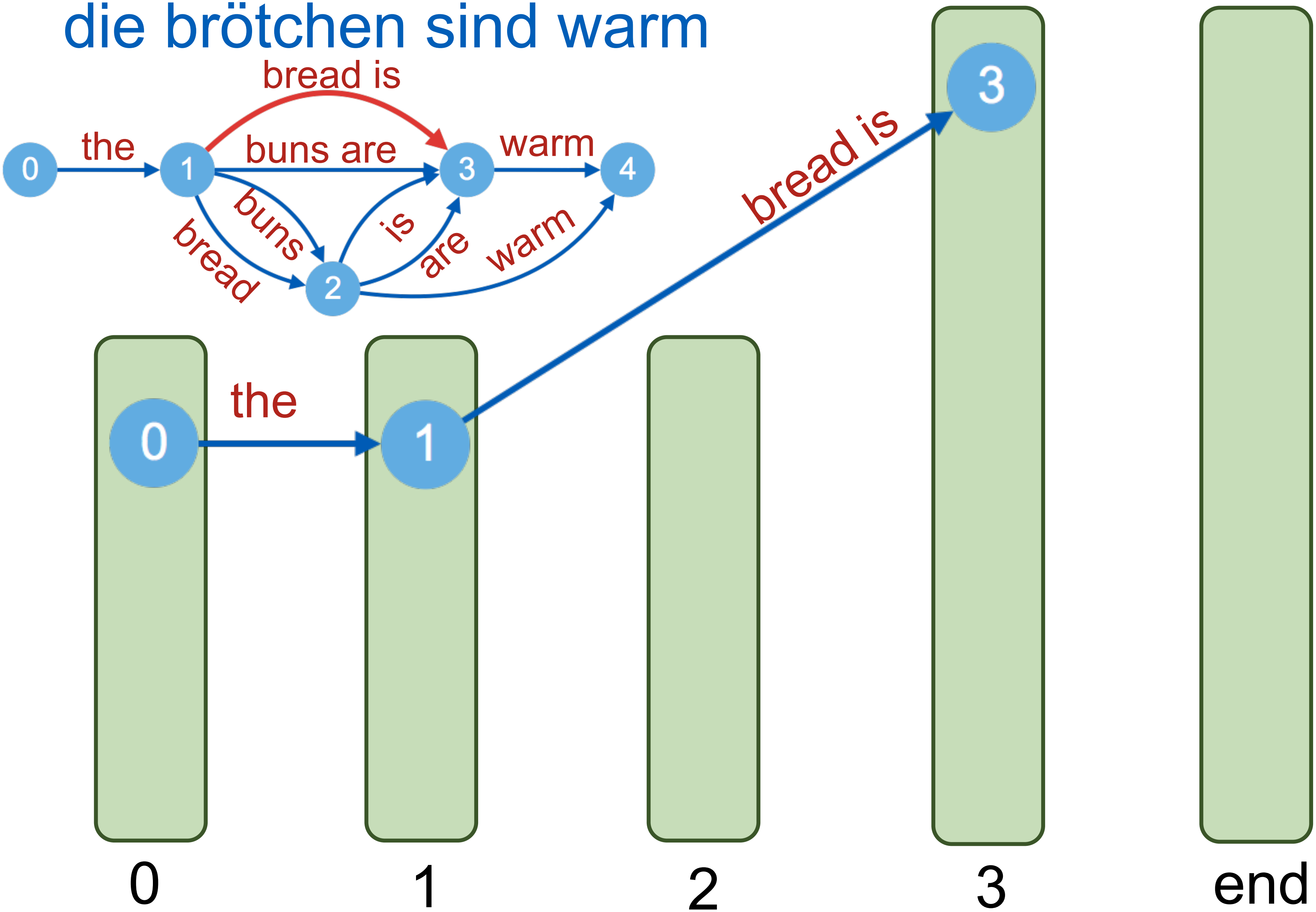
die brötchen sind warm



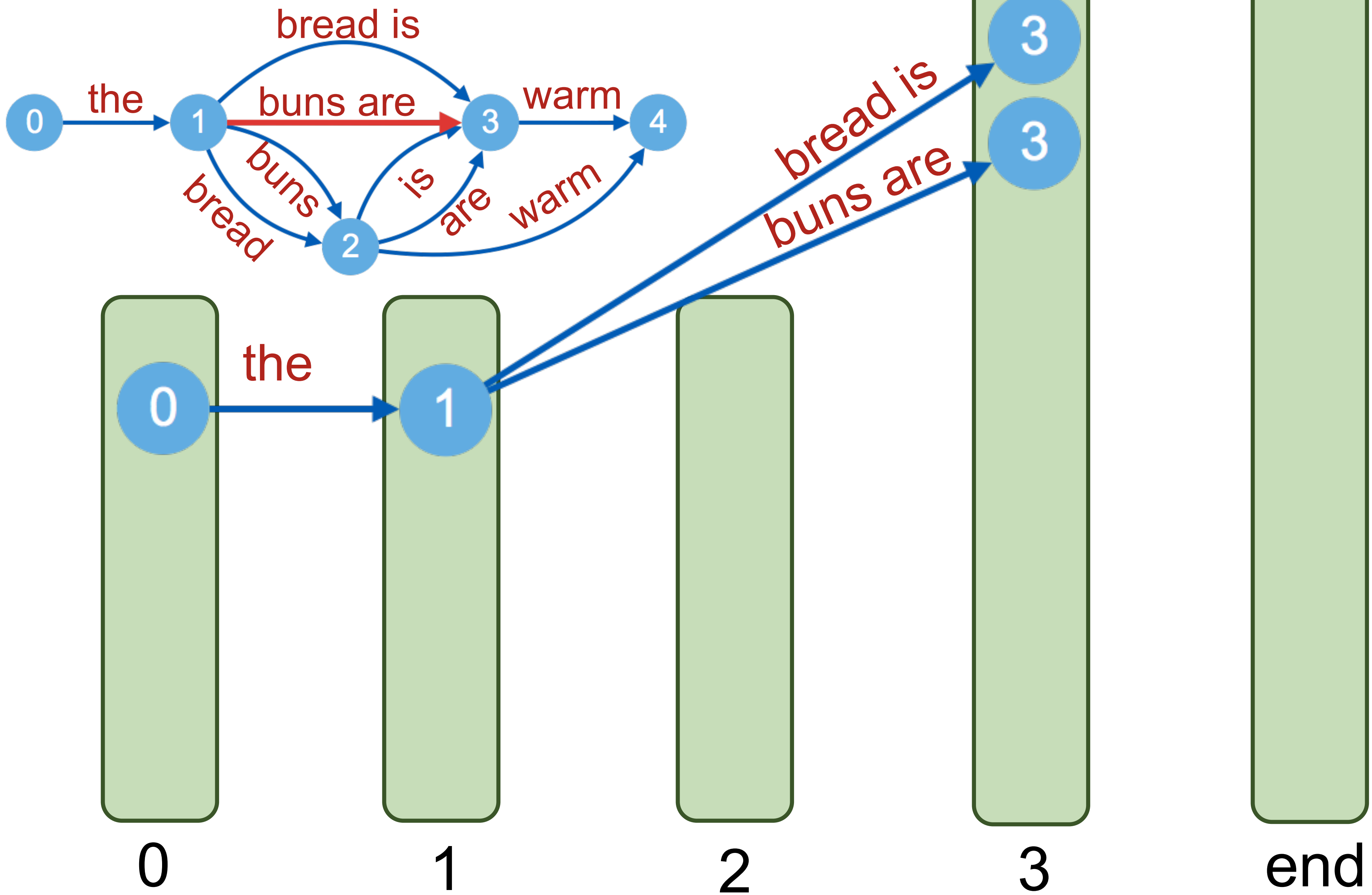
die brötchen sind warm



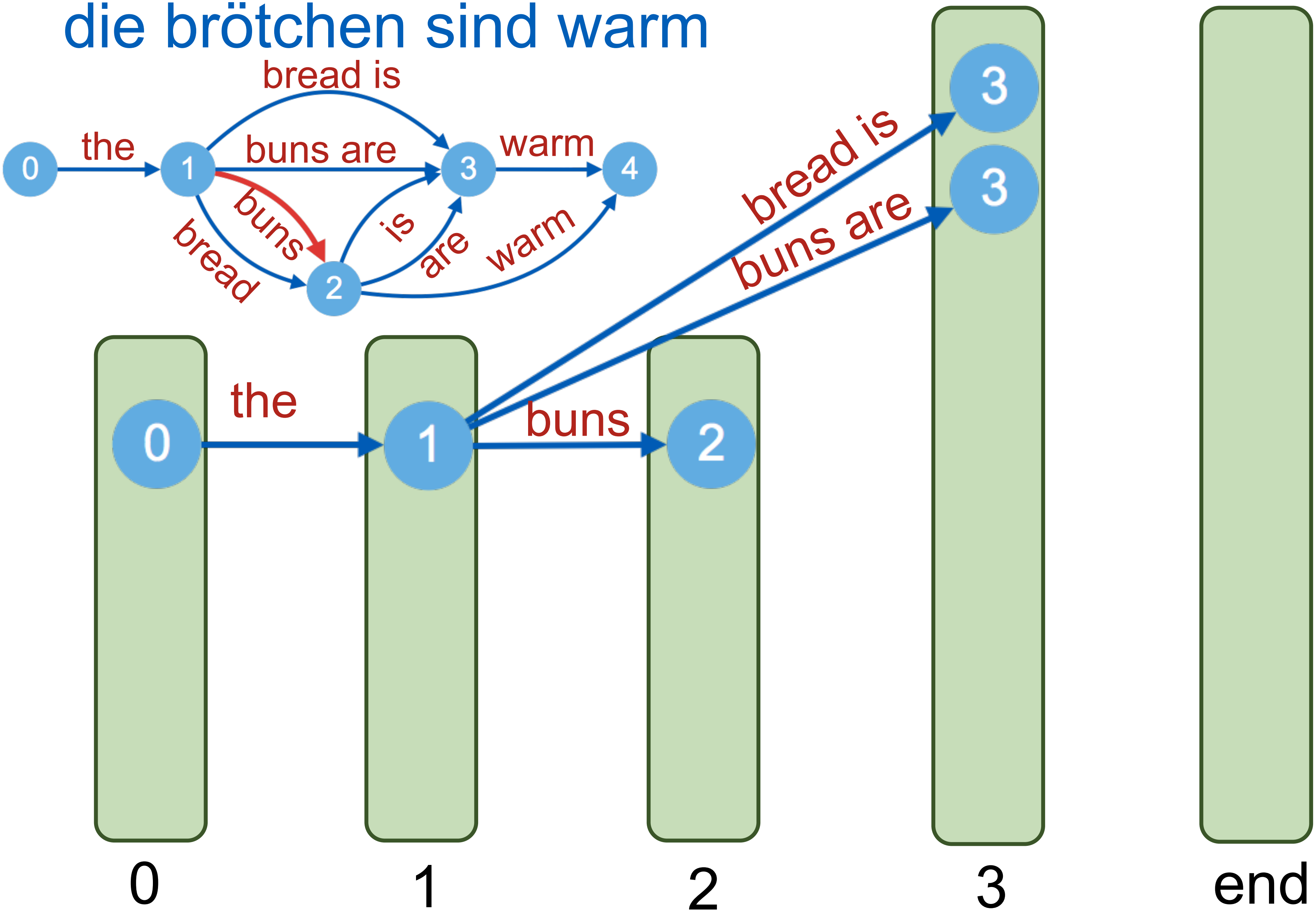
die brötchen sind warm

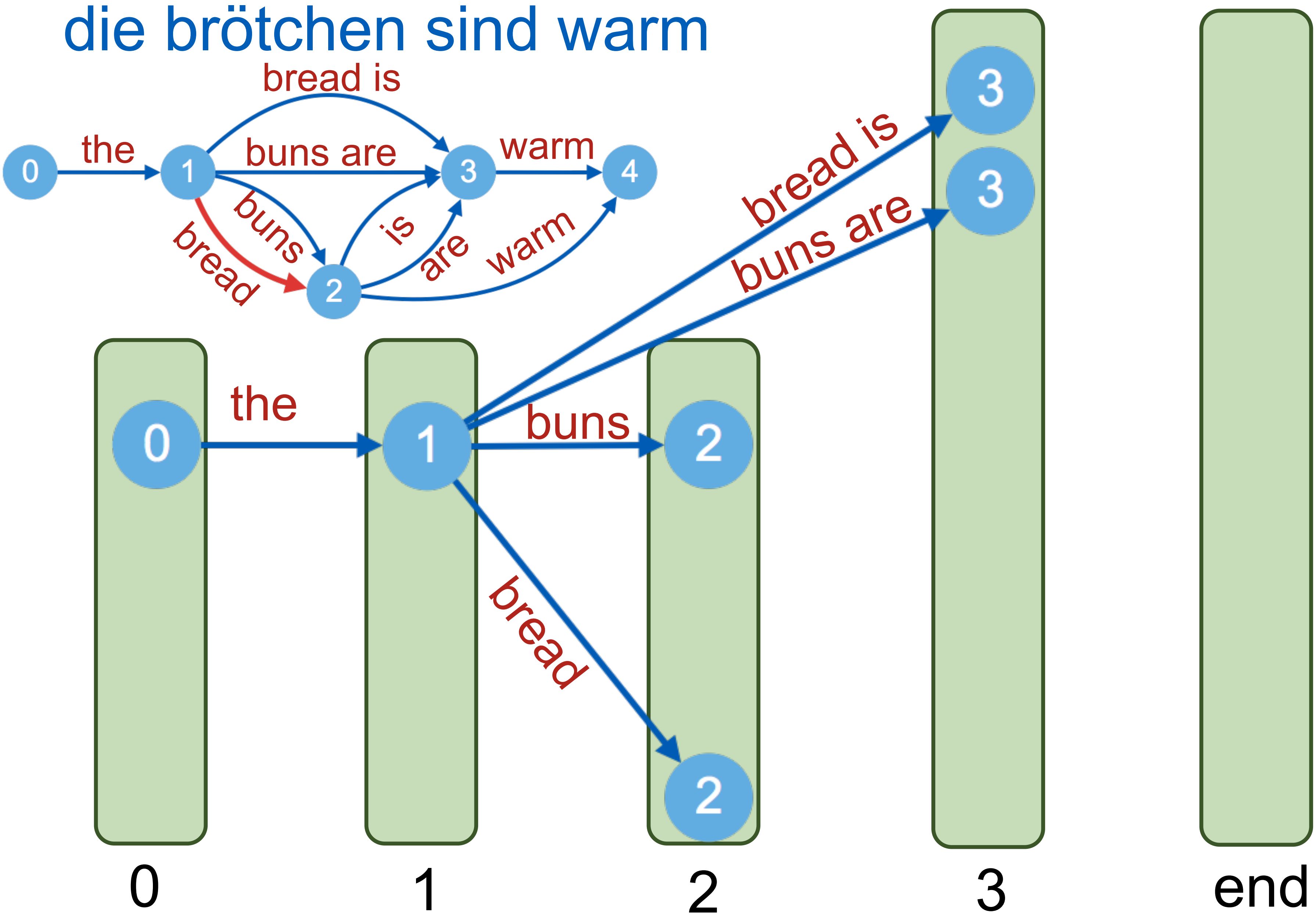


die brötchen sind warm

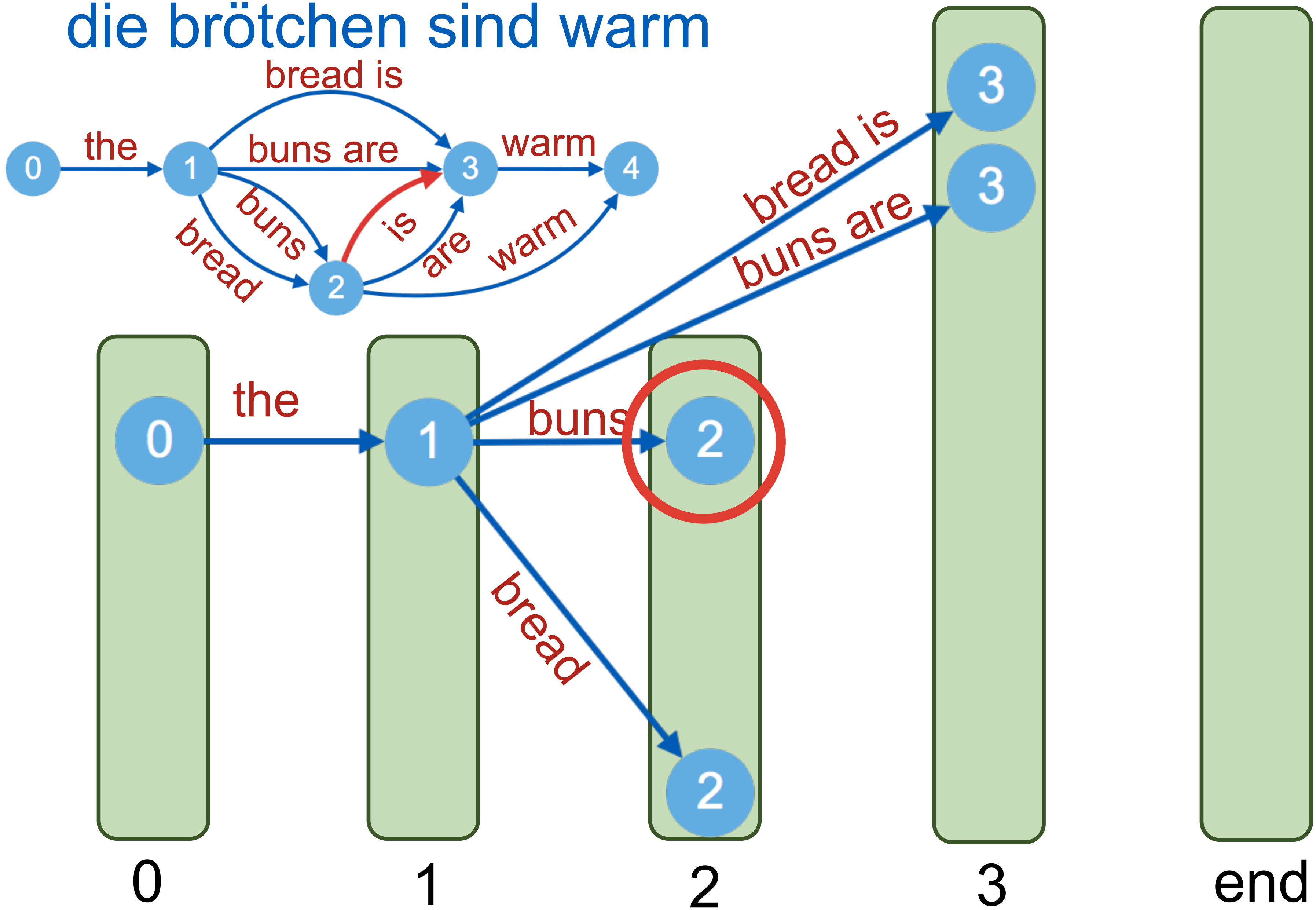


die brötchen sind warm

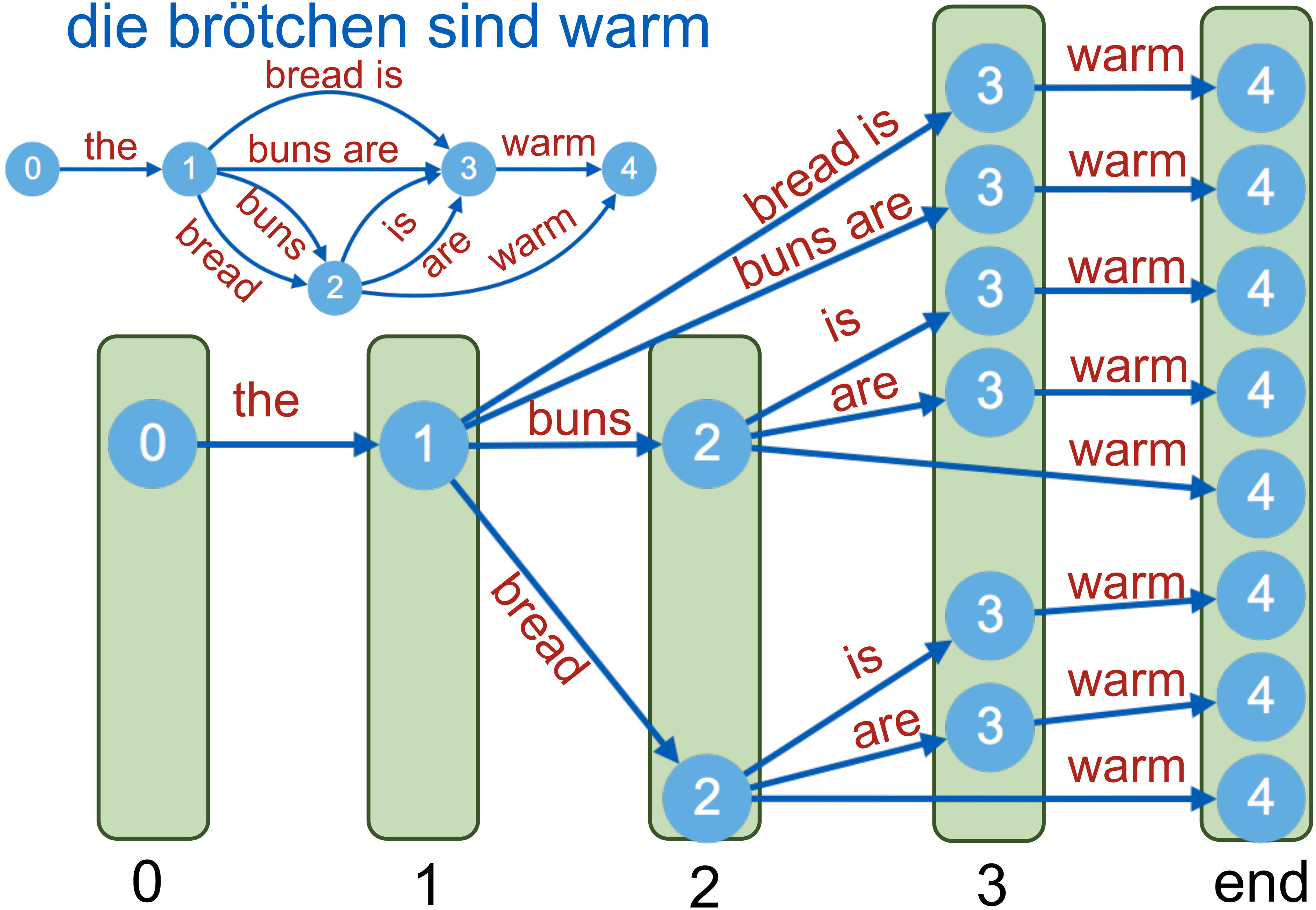



die brötchen sind warm
bread is
the
buns are
warm
buns
bread
is
are
warm
0
1
2
3
4
the
buns
bread
bread is
buns are
0
1
2
2
3
3
0
1
2
3
end

die brötchen sind warm



die brötchen sind warm



Experiments

Setting: Domain adaptation

Small **in-domain**

IT, Medical, Koran, Subtitles

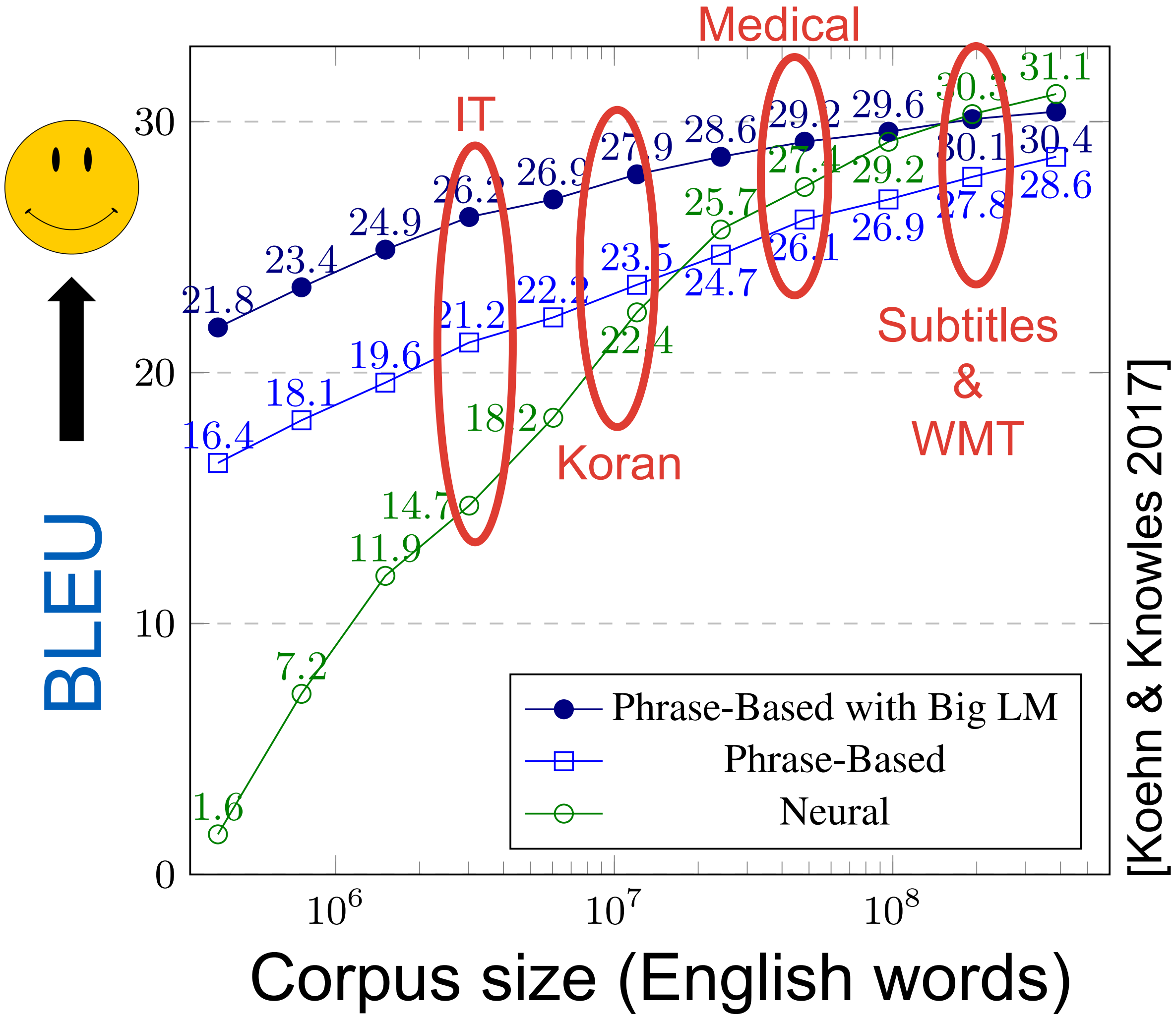
PBMT outperforms NMT

Large **out-of-domain**

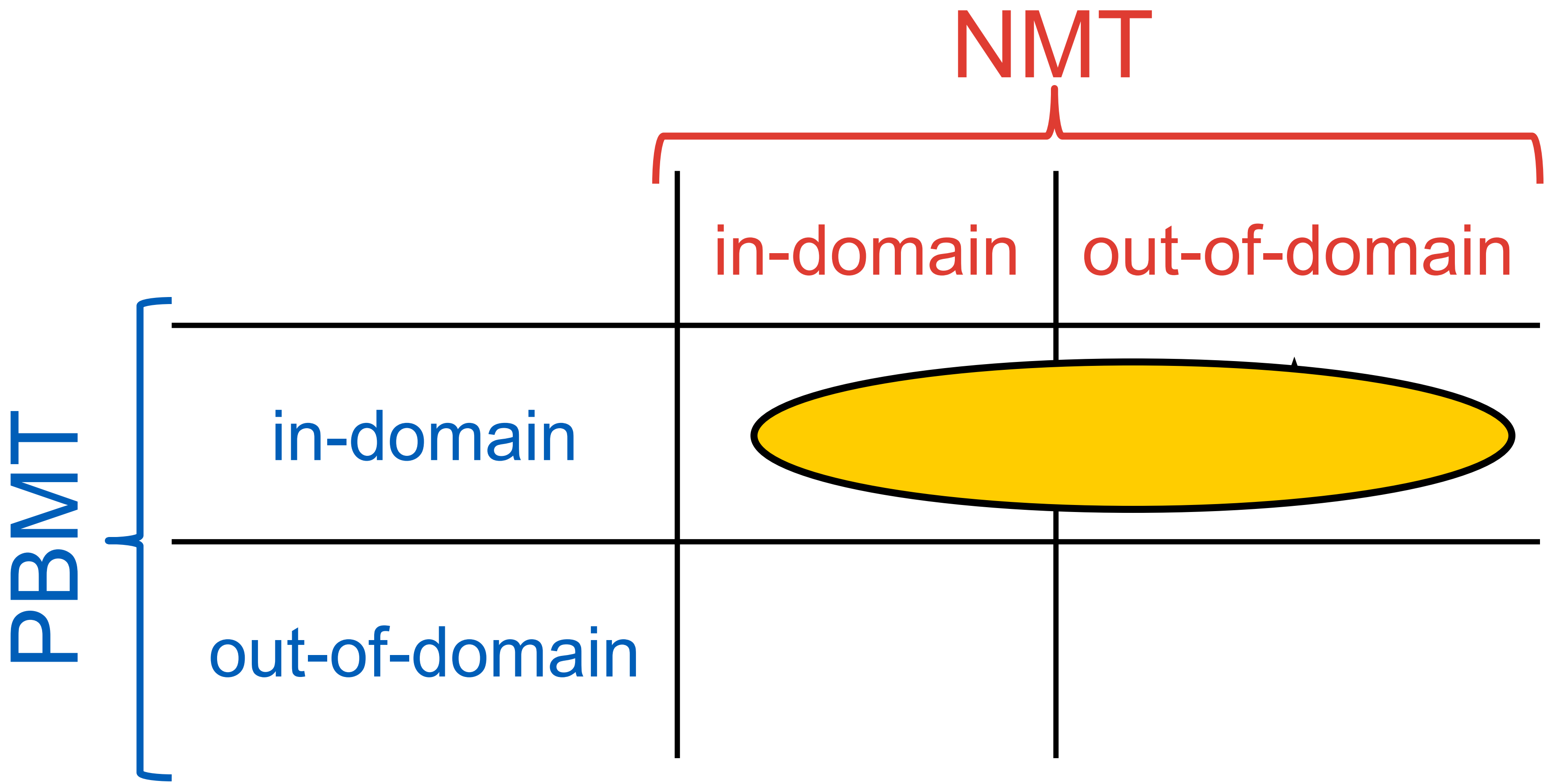
parliamentary proceedings (WMT)

NMT outperforms PBMT

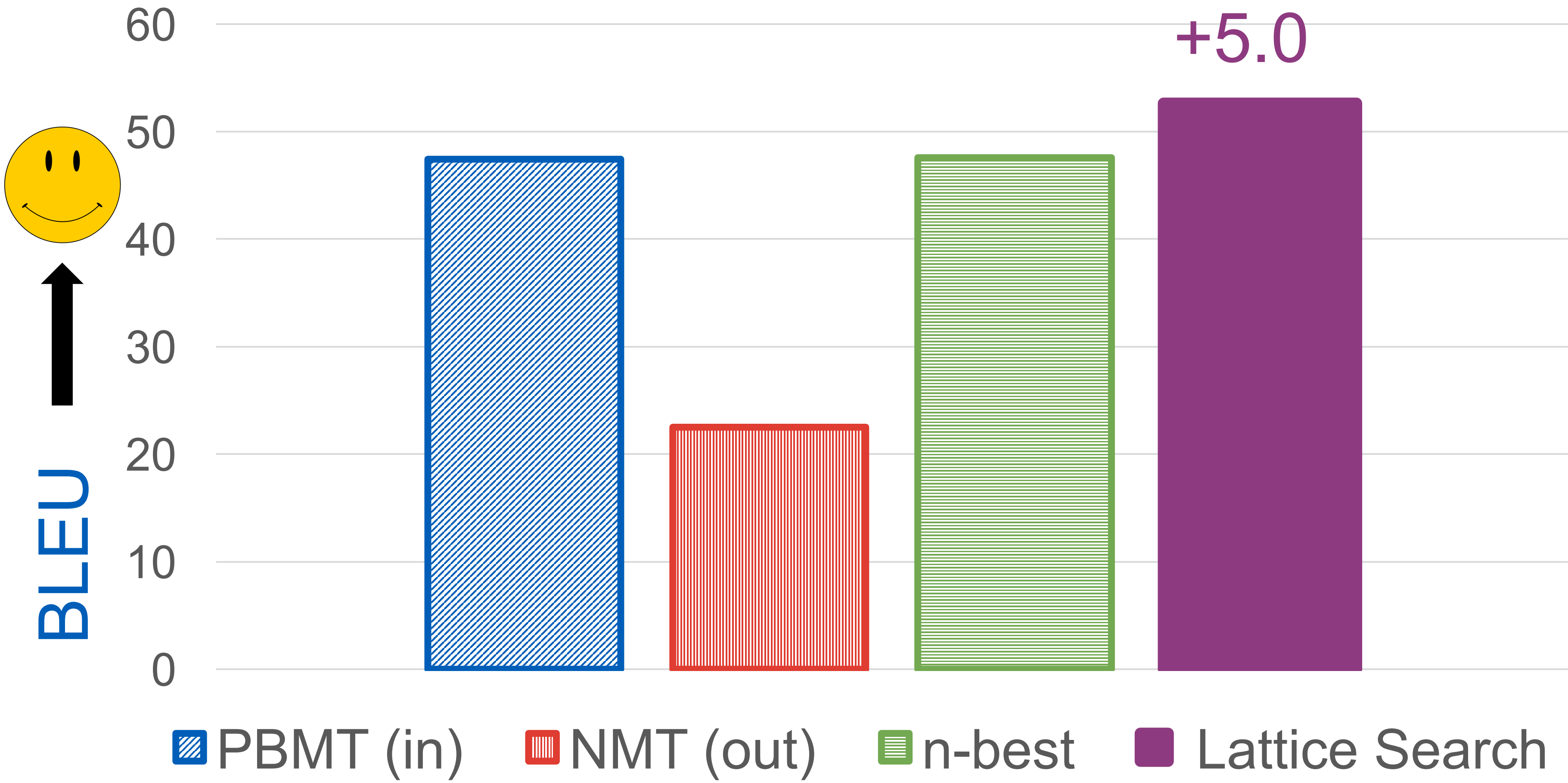
How much text do we have?



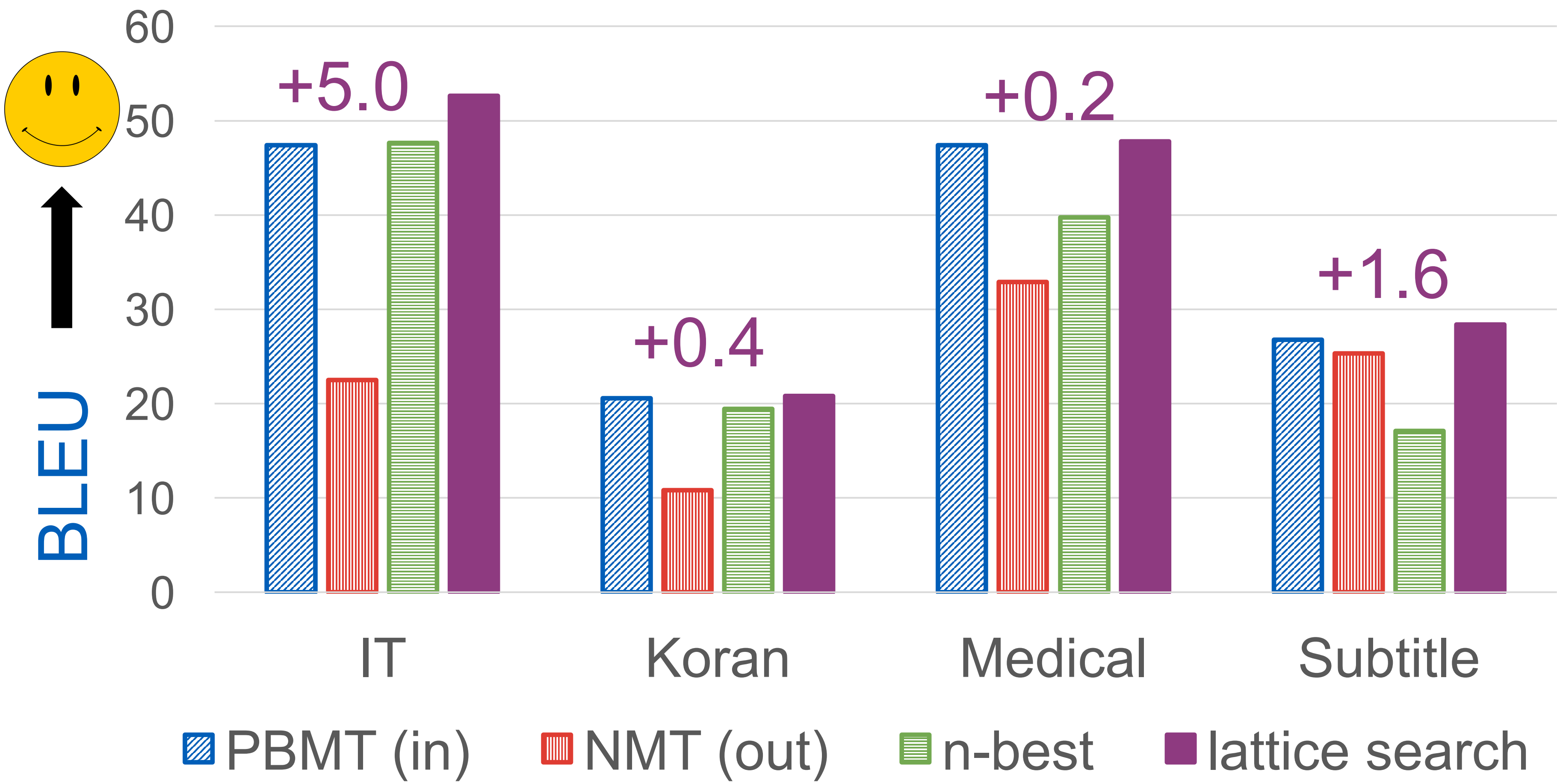
Setting: Domain adaptation



IT Results



Results


BLEU
60
50
40
30
20
10
0
+5.0
+0.4
+0.2
+1.6
IT
Koran
Medical
Subtitle
PBMT (in)
NMT (out)
n-best
lattice search

Conclusion

- Lattice search > *n*-best rescoring
- Use in-domain PBMT to constrain search space
- NMT can be in- or out-of-domain

Code:

github.com/khayrallah/nematus-lattice-search

Thanks!

This material is based upon work supported in part by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR0011-15-C-0113.

Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the Defense Advanced Research Projects Agency (DARPA).

Neural Lattice Search for Domain Adaptation in Machine Translation

Huda Khayrallah, Gaurav Kumar
Kevin Duh, Matt Post, Philipp Koehn
{huda, gkumar, kevinduh, post, phi}@cs.jhu.edu

code:
github.com/khayrallah/nematus-lattice-search

BLEU

$$min\left(1, \frac{output\ length}{reference\ length}\right)\left(\prod_{i=1}^{4} precision_i\right)^{\frac{1}{4}}$$

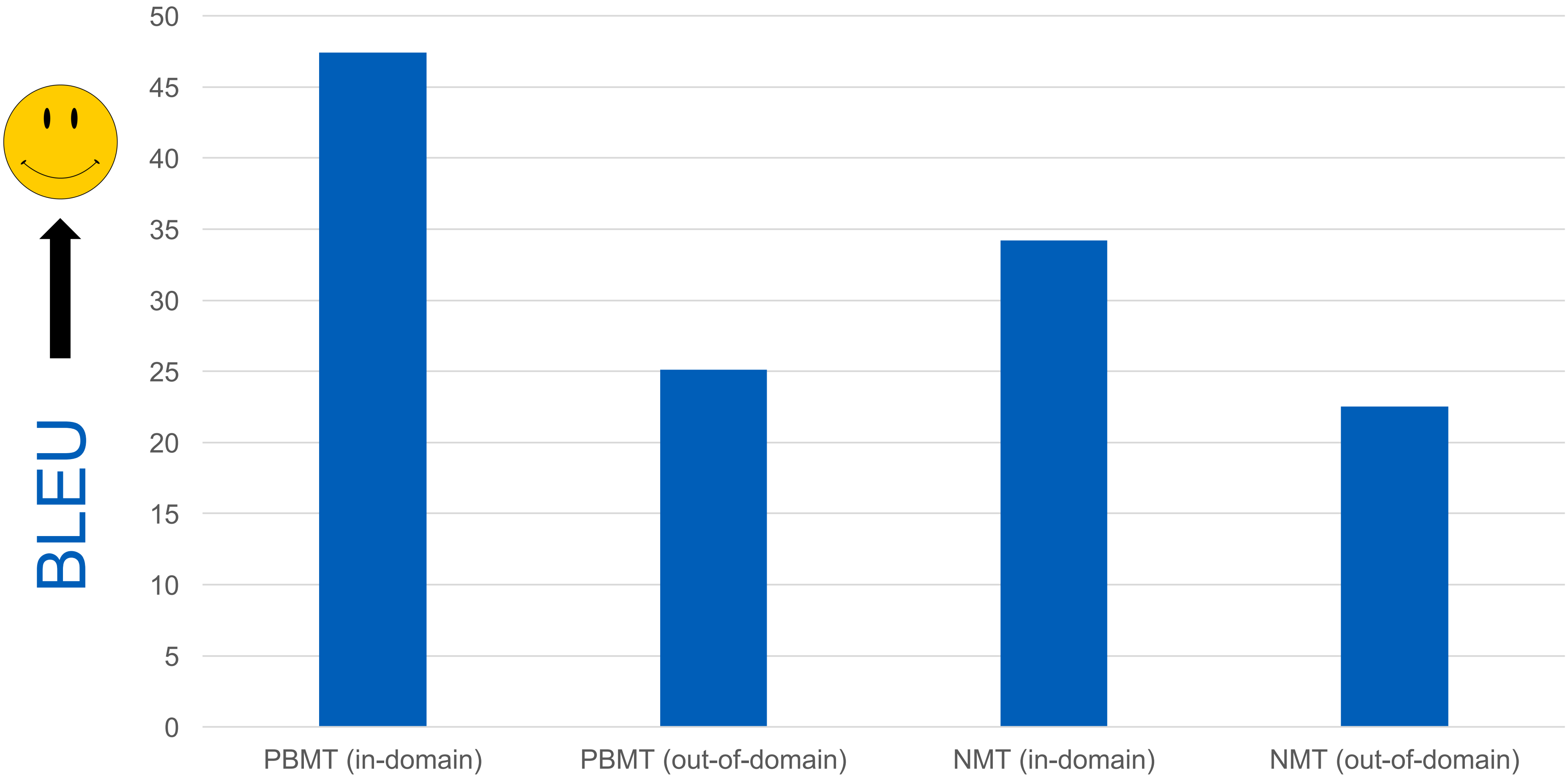
Corpus Sizes

Corpus	Words	Sentences	W/S
Medical	14,301,472	1,104,752	13
IT	3,041,677	337,817	9
Koran	9,848,539	480,421	21
Subtitles	114,371,754	13,873,398	8
EuroParl	113,165,079	4,562,102	25

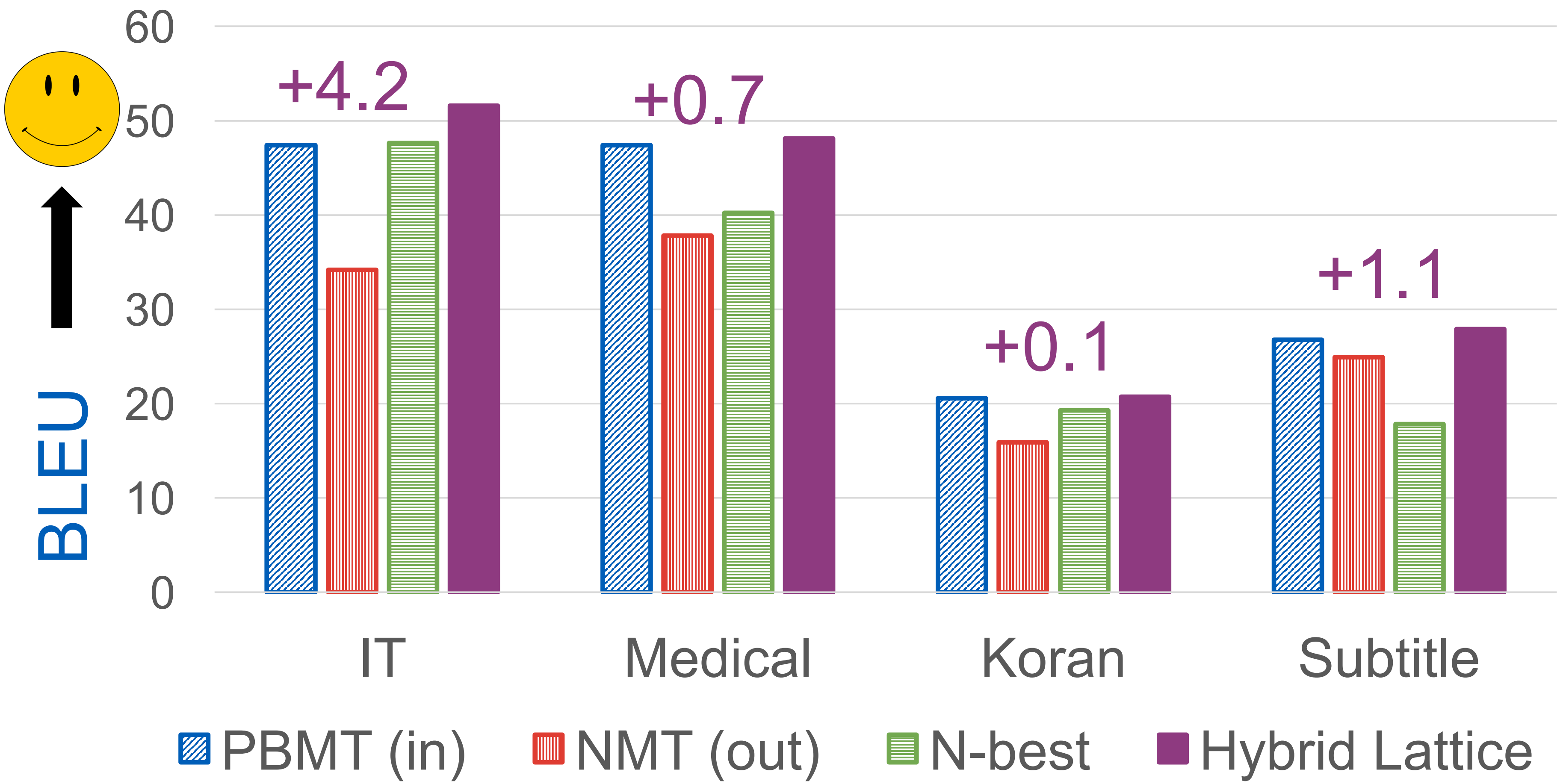
Test Domain	Training Configuration	PBMT 1-best	NMT Standard Search	N-best Rescoring	NMT Lattice Search
IT	$\text{PBMT}_{out} \times \text{NMT}_{out}$	25.1 (-0.3)	22.5 (-2.9)	22.2 (-3.2)	25.4
	$\text{PBMT}_{in} \times \text{NMT}_{in}$	47.4 (-4.2)	34.2 (-17.4)	47.6 (-4.0)	51.6
	$\text{PBMT}_{in} \times \text{NMT}_{out}$	47.4 (-5.2)	22.5 (-30.1)	47.6 (-5.0)	**52.6***
	$\text{PBMT}_{out} \times \text{NMT}_{in}$	25.1 (-2.2)	34.2 (6.9)	22.4 (-4.9)	27.3
Medical	$\text{PBMT}_{out} \times \text{NMT}_{out}$	33.3 (-0.9)	32.9 (-1.3)	30.8 (-3.4)	34.2
	$\text{PBMT}_{in} \times \text{NMT}_{in}$	47.4 (-0.7)	37.8 (-10.3)	40.2 (-7.9)	**48.1***
	$\text{PBMT}_{in} \times \text{NMT}_{out}$	47.4 (-0.4)	32.9 (-14.9)	39.7 (-8.1)	**47.8**
	$\text{PBMT}_{out} \times \text{NMT}_{in}$	33.3 (-2.7)	37.8 (1.8)	31.2 (-4.8)	36.0
Koran	$\text{PBMT}_{out} \times \text{NMT}_{out}$	14.7 (-0.2)	10.8 (-4.1)	13.9 (-1.0)	14.9
	$\text{PBMT}_{in} \times \text{NMT}_{in}$	**20.6 (-0.1)**	15.9 (-4.8)	19.3 (-1.4)	**20.7**
	$\text{PBMT}_{in} \times \text{NMT}_{out}$	**20.6 (-0.2)**	10.8 (-10.0)	19.4 (-1.4)	**20.8***
	$\text{PBMT}_{out} \times \text{NMT}_{in}$	14.7 (-1.4)	15.9 (-0.2)	13.9 (-2.2)	16.1
Subtitle	$\text{PBMT}_{out} \times \text{NMT}_{out}$	26.6 (-0.9)	25.3 (-2.2)	19.7 (-7.8)	27.5
	$\text{PBMT}_{in} \times \text{NMT}_{in}$	26.8 (-1.1)	24.9 (-3.0)	17.8 (-10.1)	**27.9**
	$\text{PBMT}_{in} \times \text{NMT}_{out}$	26.8 (-1.6)	25.3 (-3.1)	17.1 (-11.3)	**28.4***
	$\text{PBMT}_{out} \times \text{NMT}_{in}$	26.6 (-1.0)	24.9 (-2.7)	19.8 (-7.8)	27.6

Source	**Versionsinformationen ausgeben und beenden**
Reference	**output version information and exit**
PBMT	Spend version information and end
NMT	Spend and end versionary information
lattice	Print version information and exit

IT Baselines



Results

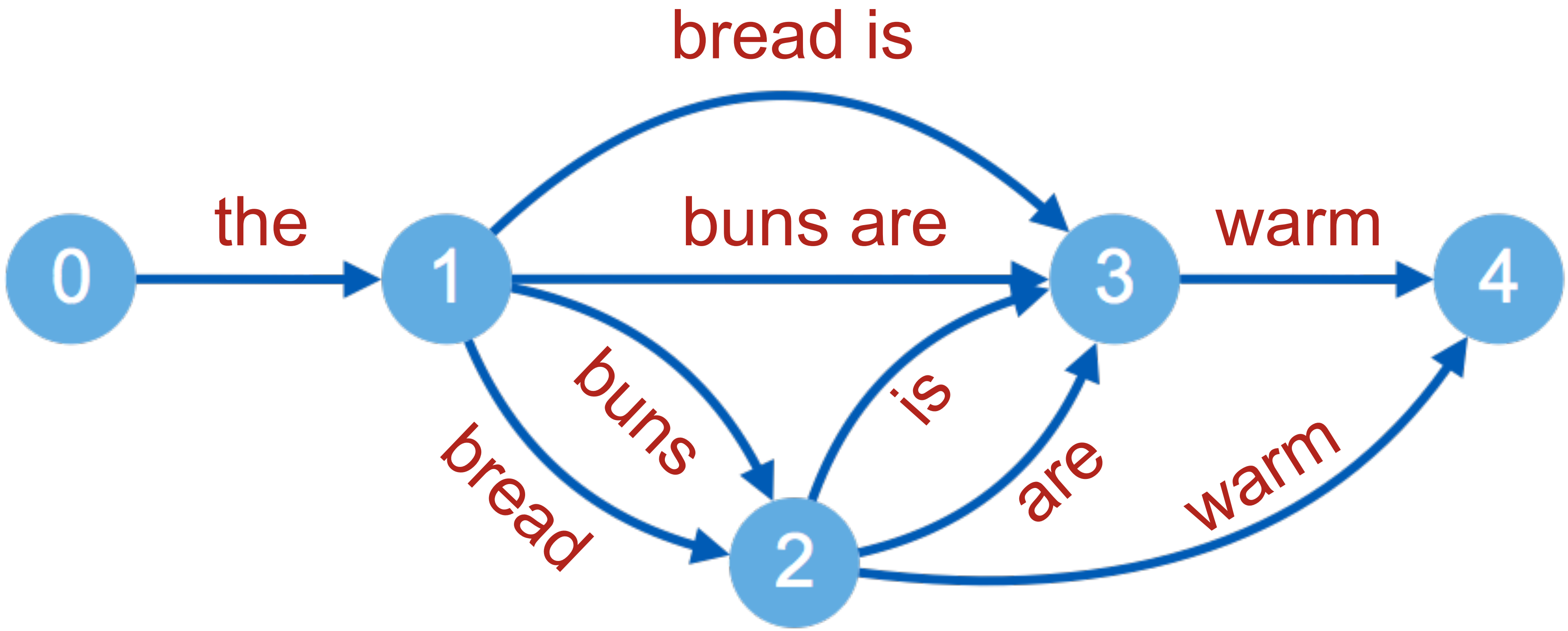


60
50
40
30
20
10
0
BLEU
+4.2
+0.7
+0.1
+1.1
IT
Medical
Koran
Subtitle
PBMT (in)
NMT (out)
N-best
Hybrid Lattice

Stack Based Decoding

- Stacks based on number of target words translated
- Keep track of:
  - Score
  - Current lattice node
  - Current neural state
  - incoming arc
  - length

die brötchen sind warm *the buns are warm*



Phrase-Based MT

$$P(target|source) = P(source|target)\,P(target)$$