

---

# A Stack-based Algorithm for Neural Lattice Rescoring

Gaurav Kumar  
Center for Language and Speech Processing  
Johns Hopkins University  
*gkumar@cs.jhu.edu*

2017/04/11



- Given a source sentence  $\mathbf{f}$ , we want to find the most likely translation  $\mathbf{e}^*$

$$e^* = \arg \max_{\mathbf{e}} p(\mathbf{e}|\mathbf{f})$$

$$= \arg \max_{\mathbf{e}} p(\mathbf{f}|\mathbf{e}) p(\mathbf{e}) \quad (\text{Bayes Rule})$$

$$= \arg \max_{\mathbf{e}} \sum_{\mathbf{a}} p(\mathbf{f}, \mathbf{a}|\mathbf{e}) p(\mathbf{e}) \quad (\text{Marginalize over alignments})$$

- The alignments  $\mathbf{a}$  are latent.  $p(\mathbf{f}, \mathbf{a}|\mathbf{e})$  is typically decomposed as:
  - Lexical/Phrase **Translation Model**
  - An **Alignment/Distortion Model**
- $p(\mathbf{e})$  is the **Language Model**

- Decoding may find features besides the ones derived from the generative model useful
  - reordering (distortion) model
  - phrase/word translation model
  - language models
  - word count
  - phrase count
- The use of multiple features typically takes the form of a log-linear model

$$p(\mathbf{e}|\mathbf{f}) = \frac{\sum_i \lambda_i f_i}{Z} \quad (Z \text{ is the partition function})$$

Where each “feature”  $f_i$  is exponentially scaled by a weight  $\lambda_i$   
Features are not necessarily valid probabilities

# Learning to align and translate

Joint learning of alignment and translation (*Bahdanau et al., 2015*)

- One model for translation and alignment
- Extends the standard RNN encoder-decoder framework for neural network based machine translation
- Allows the use of an alignment based soft search over the input

# RNN encoder-decoder

- **Encoder** : Given any sequence of vectors  $(f_1, \dots, f_J)$

$$s_j = r(f_j, s_{j-1}) \quad (\text{Hidden state})$$

$$c = q(\{s_1, \dots, s_J\}) \quad (\text{The context vector})$$

where  $s_j \in \mathbb{R}^n$  is the hidden state at time  $j$ ,  $c$  is the context vector generated from the hidden states and  $r$  and  $q$  are some non-linear functions.

- **Decoder** : Predict  $e_i$  given  $e_1, \dots, e_{i-1}$  and the context  $c$ .

$$p(\mathbf{e}) = \prod_{i=1}^I p(e_i | \{e_1, \dots, e_{i-1}\}, c) \quad (\text{Joint probability})$$

$$p(e_t | \{e_1, \dots, e_{i-1}\}, c) = g(e_{i-1}, t_i, c) \quad (\text{Conditional probability})$$

where  $t_i$  is the hidden state of the RNN and  $g$  is some non-linear function that outputs a probability.

# Neural Machine Translation

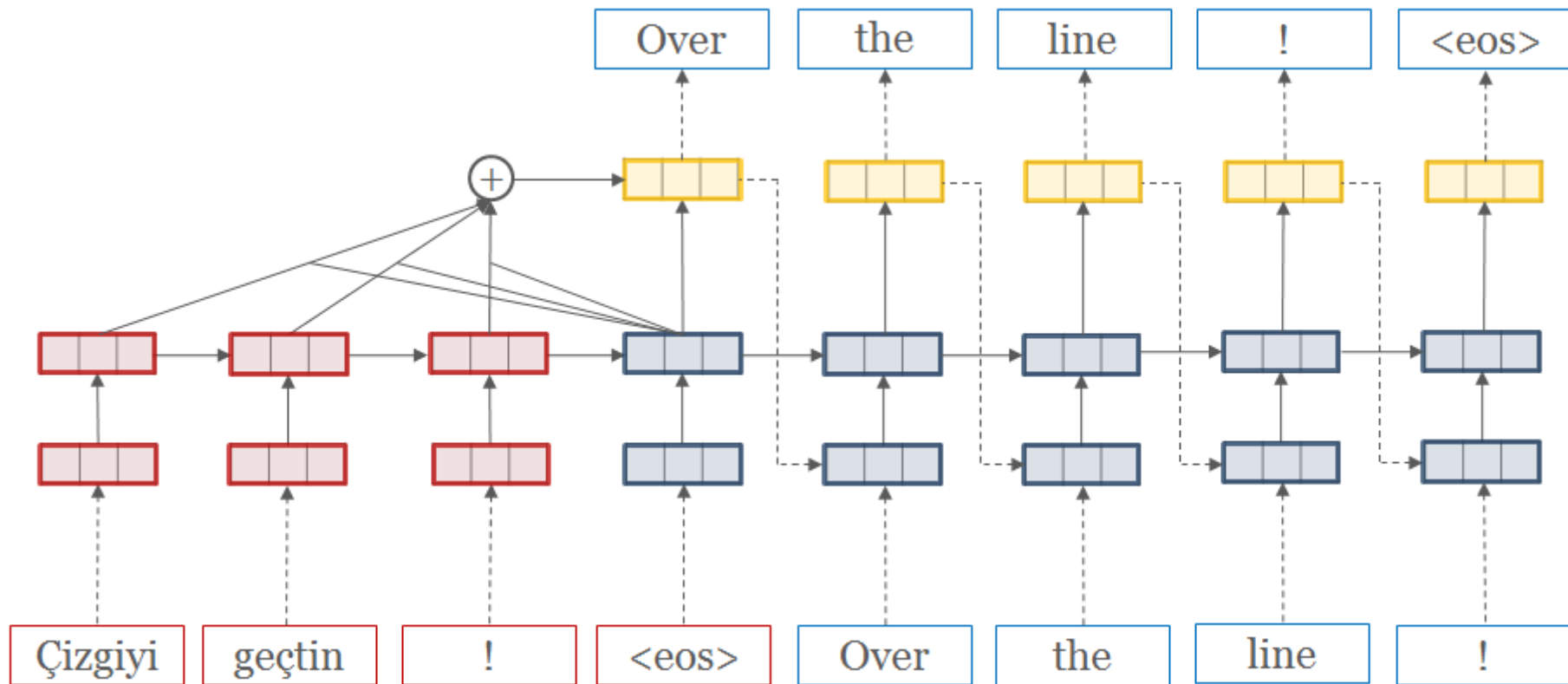


Figure 1: Neural Machine Translation with attention (Image from opennmt.net)

# Neural Machine Translation in 2015







		output language				
input language	Czech 		jhu-smt Moses			
		German 	Matthias Huck uedin- pbt-			
	obo chimera	UMontreal- MILA Neural	English 	atoral abumatran-	Benjamin Marie LIMSI- CNRS	lexi2 Edinburgh
			barry uedin- pbt-	Finnish 		
			Benjamin Marie LIMSI- CNRS		French 	
			jhu-smt Moses			Russian 

Figure 2: WMT2015 evaluation results for language-pairs (Image from [matrix.statmt.org](http://matrix.statmt.org))

# Neural Machine Translation in 2016










		output language					
i n p u t  l a n g u a g e	<b>Czech</b> 		barry uedin-nmt				
		<b>German</b> 	rsennrich uedin-nmt-				
	barry uedin- nmt	rsennrich uedin- nmt-	<b>English</b> 	atoral abumatran-	qt21 QT21_HimL_	rsennrich uedin- nmt-	post jhu- hltcoe
			barry uedin-pbmt	<b>Finnish</b> 			
			barry uedin-pbmt		<b>Romanian</b> 		
			Marcin Junczys- Dowmunt AMU- UEDIN			<b>Russian</b> 	
			emrebektas tbtk-sysco				<b>Turkish</b> 

Figure 3: WMT2016 evaluation results for language-pairs (Image from [matrix.statmt.org](http://matrix.statmt.org))



# Are we done?



8

- As more and more parallel data becomes available, the performance of the NMT systems is only going to improve.
- Research into using monolingual data is already proving successful (TODO: citation here).
- More complex encoder-decoder models are being proposed every week.
- Hardware scaling helps supports more parameters and more complex models.

**When does NMT not perform well?**

# NMT Challenges : Low Resource



Language	Train size	Test size	SBMT BLEU	NMT BLEU
Hausa	1.0m	11.3K	23.7	16.8
Turkish	1.4m	11.6K	20.4	11.4
Uzbek	1.8m	11.5K	17.9	10.7
Urdu	0.2m	11.4K	17.9	5.2

Figure 4: Performance of NMT models vs. string-to-tree models for low resource languages (Image Zoph et al., 2016)

## Current research

- Transfer learning : Zoph et al., 2016
- Multi-way, multi-lingual NMT : Firat et al., 2016

# NMT Challenges : Out of domain



- A problem not unique to NMT
- A fundamental challenge for DARPA Lorelei
- Assume that you have access to parallel text in the following domains: religious, legal and IT. Your job is to come up with a translation system that can be used to assist and converse with earthquake victims.
- Possibly worse for NMT because of the drastically different style of writing used in the out of domain training text. This is the trouble with using source conditioned language models.

# NMT Challenges : Out of domain

System	Law	Medical	IT	Koran	Subtitles
All	-1.3	+2.9	-9.4	$\pm 0.0$	+5.6
Law	-3.3	-6.1	-3.4	-0.9	-3.2
Medical	-6.3	-4.1	-6.5	-1.4	-4.4
IT	-1.8	-1.2	+2.3	+0.2	-0.8
Koran	-1.4	-2.1	-2.3	-2.9	-4.5
Subtitles	-2.9	-8.5	-4.4	+0.6	+3.8

Table 1: Relative performance of NMT systems with respect to PBMT systems for out-of-domain test sets in German-English (From Philipp Koehn)

# NMT Challenges : The UNK problem



- NMT systems do not copy words from the source into the target if an unknown word is encountered.
- For languages which have a large vocabulary size or greater morphological complexity, producing an UNK is safe
- Degenerate solution, if enough UNKs are in the training data, safely produce an UNK during translation

An example from Romanian-English (newstest2016):

**Ref** : 46 percent said they are leaving the door open to switching candidates .

**Moses** : 46 % say porti?a leaves open the possibility of changing the option .

**NMT** : 46 per cent affirmative the unk tag # selunk tag # selunk

# NMT Challenges : The rare word problem

13



**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.*)

Figure 5: A mistake made by an NMT system on a low-frequency content word  
(Image from Arthur et al., 2016)

- Rare words which belong to a common word class are often confused.
- This problem is worse for words that are of interest for downstream NLP tasks such as NER.

# NMT Challenges : The rare word problem

14



## Current Research

- Subword translation (Sennrich et al., 2015)
- Character level NMT (Ling et al., 2015)
- Incorporations of lexicons (Arthur et al., 2016)
- Tracking source words which produced OOVs (Luong et al., 2015)

# NMT Challenges : Length ratios & hallucination



Ref	ban urged the five permanent members to show the solidarity and unity they did in achieving an iran nuclear deal in addressing the syria crisis .
Moses	ban urged the five permanent members to show solidarity and unity shown when they failed to reach a deal on iran 's nuclear weapons , thus addressing the crisis in syria .
NMT	ban called on the five permanent members of the lib Dems to give pumpkins of solidarity with the arthritis unit , then the cudgel reeled it sunk nkey an agreement on iran 's nuclear weapons , to handle the crisis in syria .

Table 2: An example translation from the Romanian-English newstest2016 test set.



# NMT Challenges : Ignoring source context

16



- No explicit accountability for translating all source words with NMT models

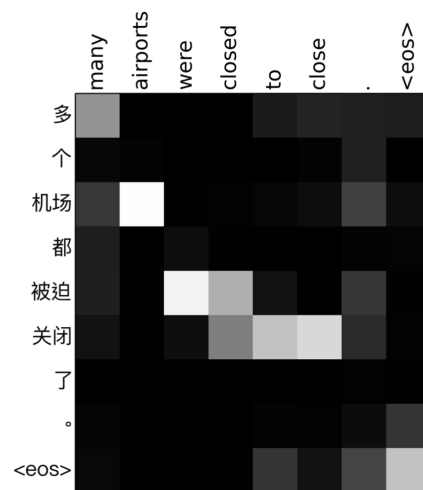


Figure 6: Ignoring source words in translation with NMT models (Image from Tu et al., 2016)

## Current Research:

- Modeling coverage vectors (Tu et al., 2016, Mi et al., 2016)
- Supervised alignments (Liu et al., 2016)

# Adequacy vs. Fluency



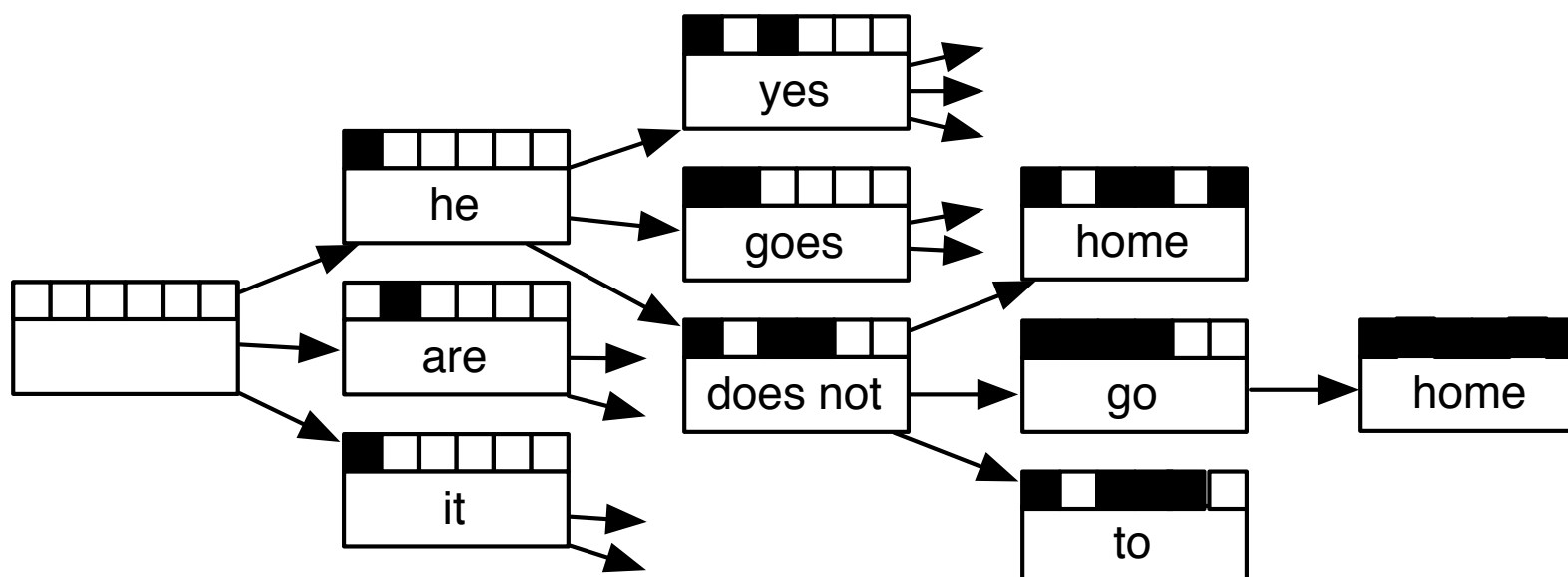
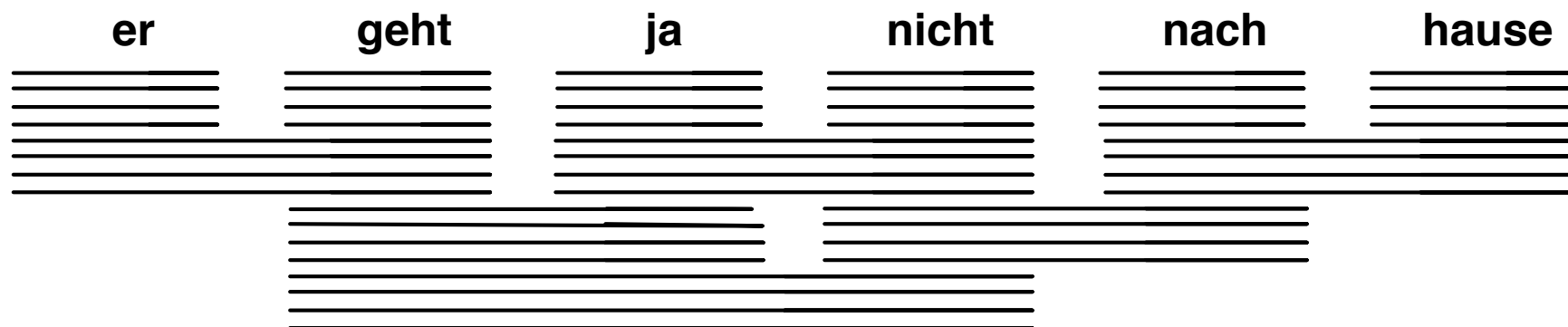
- SMT systems are tasked with the explicit translation of each component within the source sentence (**adequate**).
- NMT systems produce text which is generally fluent and fairly well conditioned on the source sentence (**fluent**).

We plan to combine these benefits by using the SMT system to **constrain the hypothesis space of adequate translations** available to the NMT system which will choose the most fluent one.

- **System combination** : Using  $n$ -best lists for combination (via features or otherwise) for multiple NMT and SMT systems if common.
- **Moses with NMT features** : Use the NMT score as a feature in PBMT (Junczys-Dowmunt et al., 2016).
- **Promoting diversity in beam search** (Vijayakumar et al., 2016)
- **Using alternate objective functions** while training NMT systems to increase diversity (Li et al., 2016)
- **Minimize Bayes risk with respect to lattices** (Stahlberg et al., 2017)

# SMT Search graphs

19



# Re-scoring SMT Search graphs

- Search graphs (which can be converted to word lattices) represent a compact and potentially diverse set of translation hypotheses.
- In comparison,  $n$ -best lists may lack this diversity.
- Search graphs also allow efficient traversal of the hypothesis space, eliminating entire sub-graphs of translations if their prefix scores are bad. This is not possible with  $n$ -best lists.
- For this study, we limit the role of the SMT system to constraining the search space. We discard all of the SMT features and scores on the lattice.
- Out-of-domain neural models are useful again, since they are choosing from a constrained adequate hypothesis space.

# Stack-based re-scoring algorithm

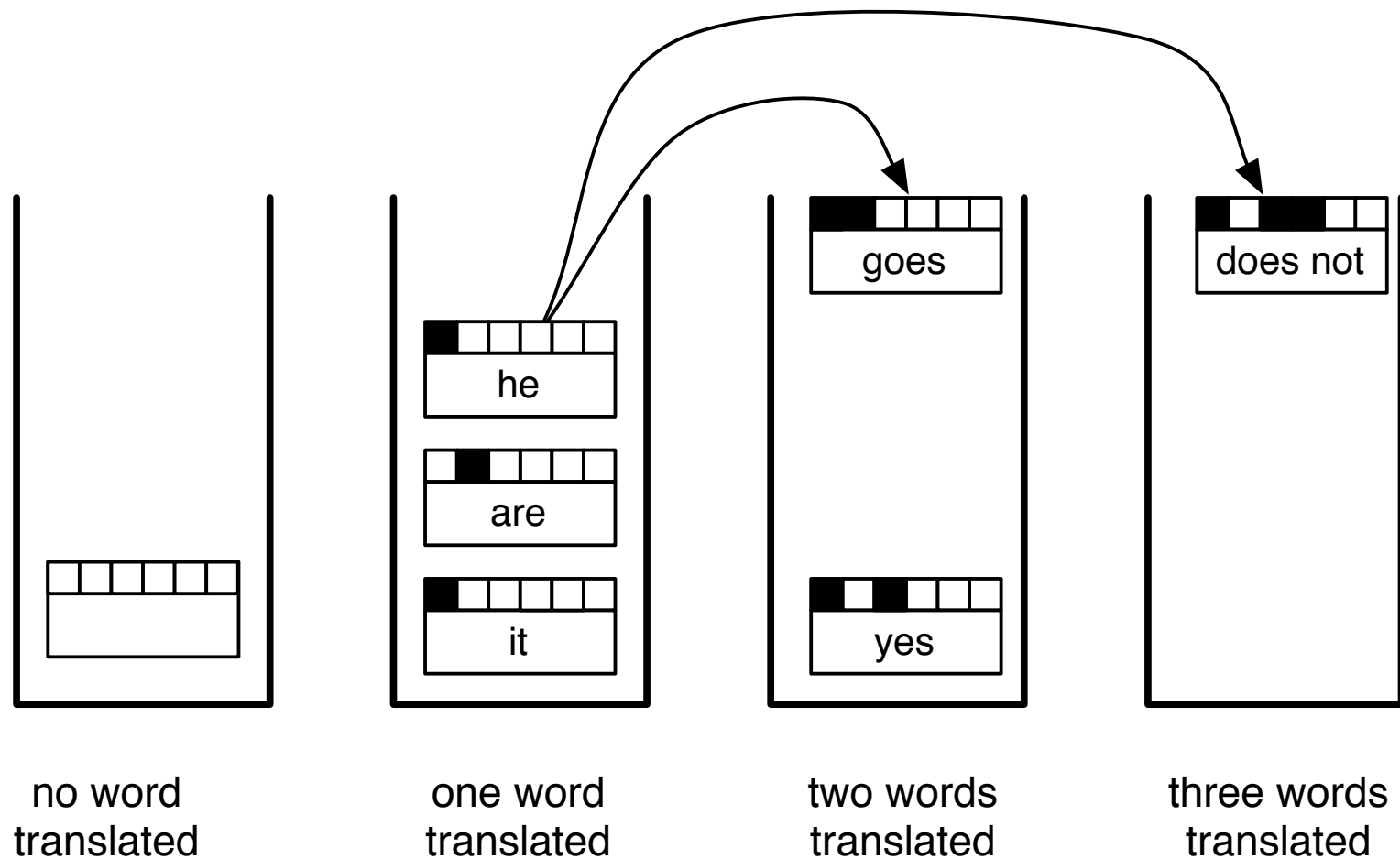


Figure 7: Phrase based stack decoding (Image from Philipp Koehn)

# Stack-based re-scoring algorithm

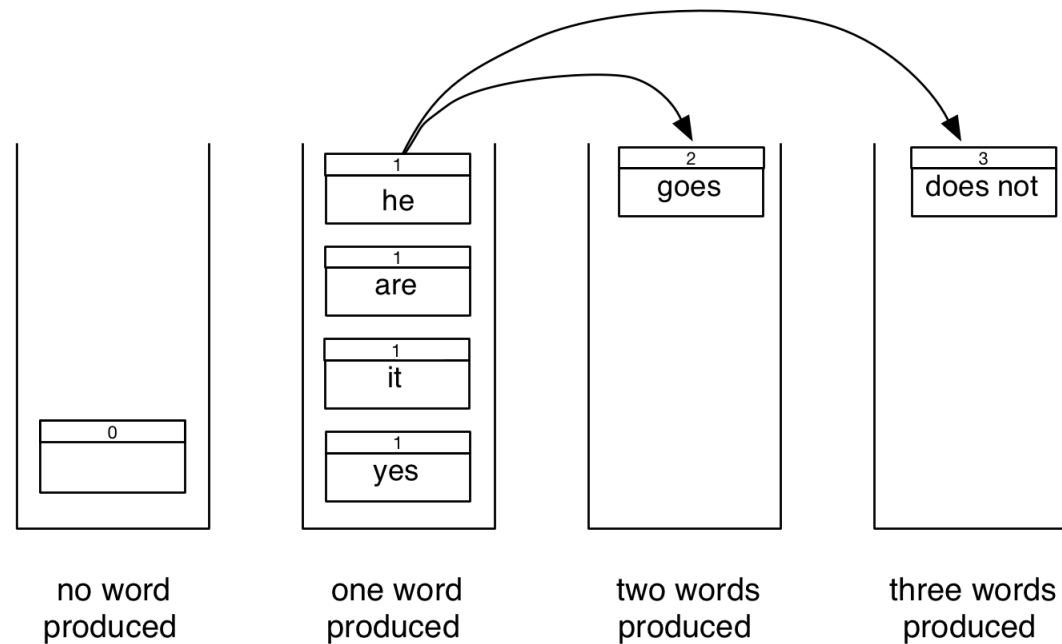


Figure 8: Search graph stack rescoring

- All complete hypotheses are moved to a “complete” stack.
- All scores are length normalized to avoid a length bias.

# Recombination

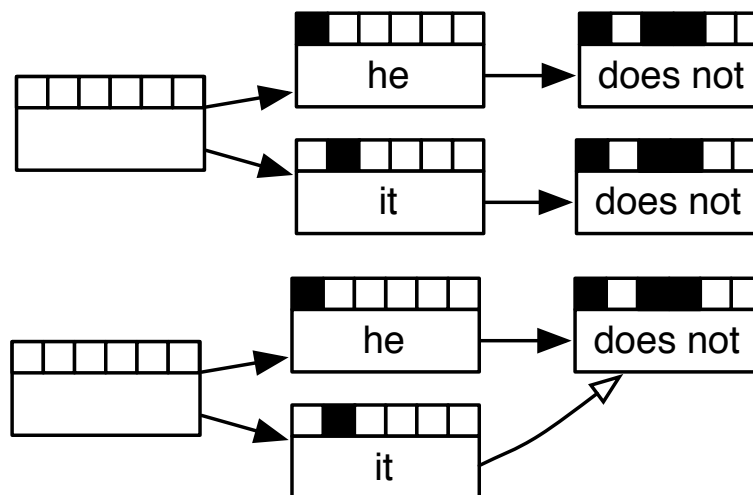


Figure 9: Recombination in a search graph when the states are indistinguishable.

- When states in a search graph are indistinguishable with respect to their contexts but have different scores.
- Drop the worse one in a traditional SMT stack decoder.
- How do we handle this with NMT re-scoring, since each path has a unique context?



For lattice re-scoring with stacks:

- Treat each state in a stack (now, an RNN state) as unique
- However, only keep the top  $k$  entries in a stack when processing it (Histogram pruning)
- Expand all possibilities from the best  $k$  entries of the stack currently being processed.

# Stack-based re-scoring algorithm

- Explores an “adequate” hypothesis space of translations with neural translation models.
- The exploration space is typically more diverse than  $n$ -best lists.
- No re-training of the models is required.
- This is faster than using the NMT feature function in SMT systems.
- When the SMT system is more robust than the NMT models, this may potentially improve quality.

# Experiments

- 

- 

Desc datasets, systems

## Results: SMT better, NMT worse

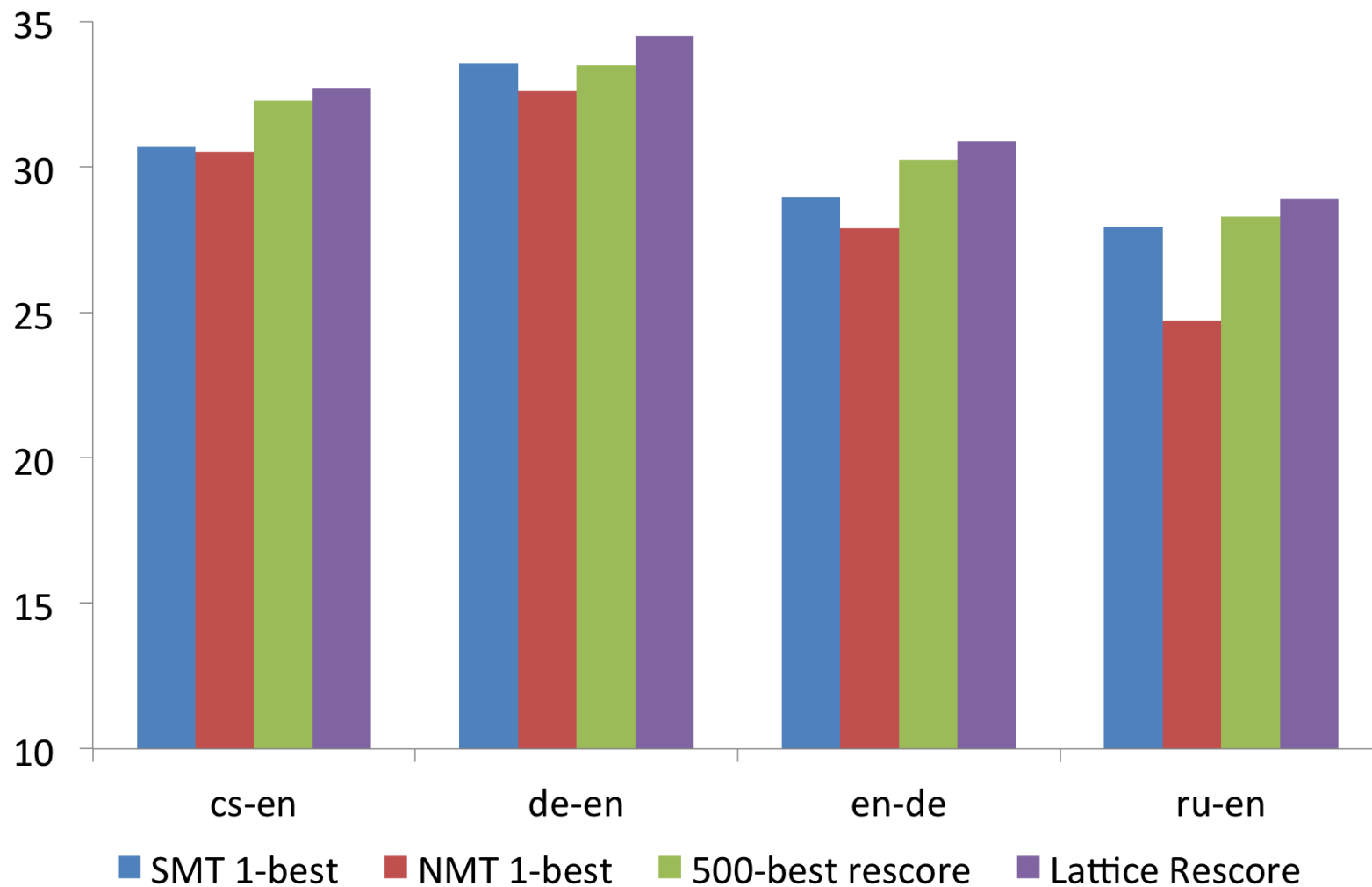


Figure 10: Lattice re-scoring performs the best when SMT is better than NMT.

## Results: SMT worse, NMT better

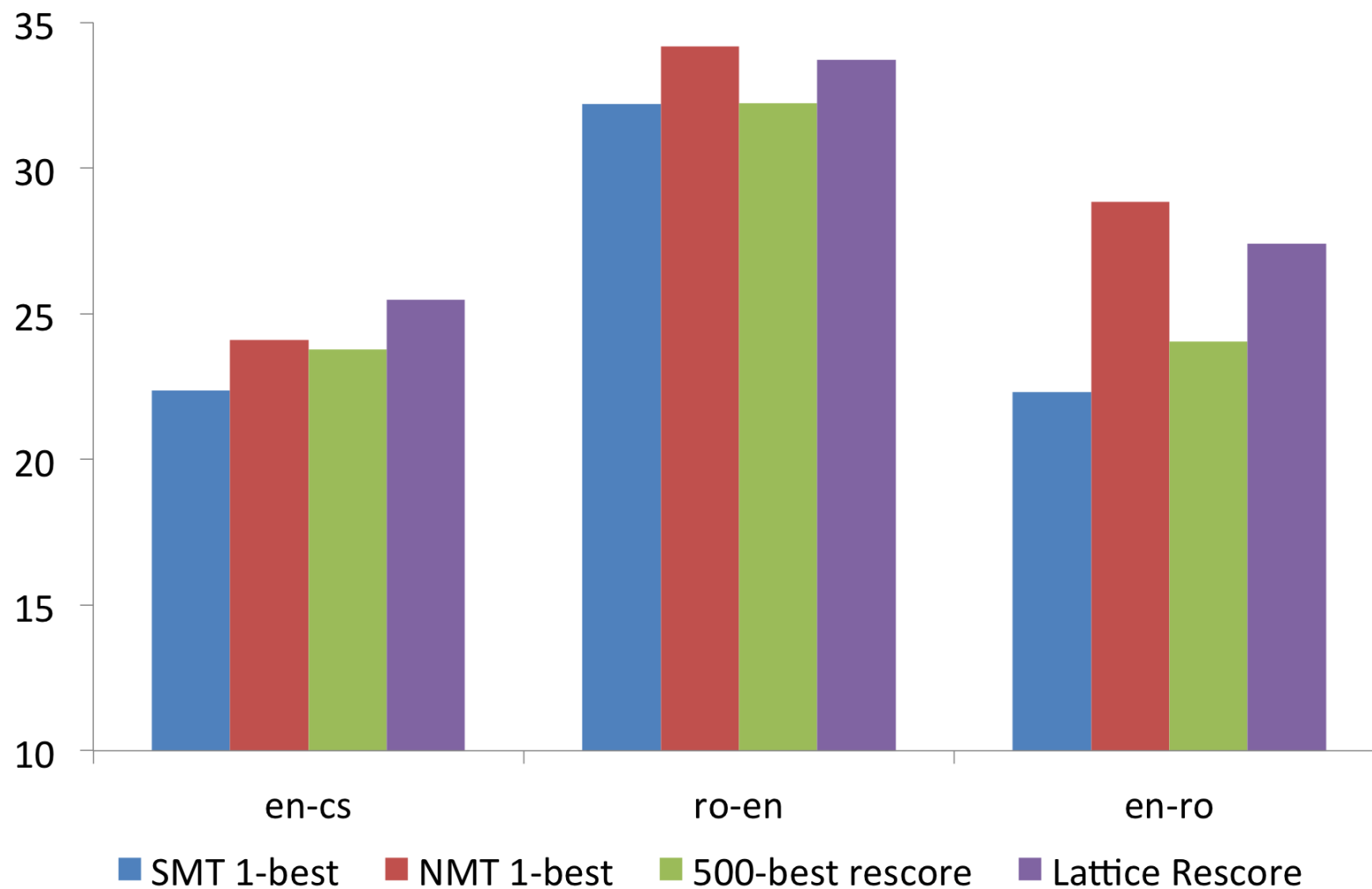


Figure 11: Lattice re-scoring performance when NMT is better than SMT.

# The effect of pruning

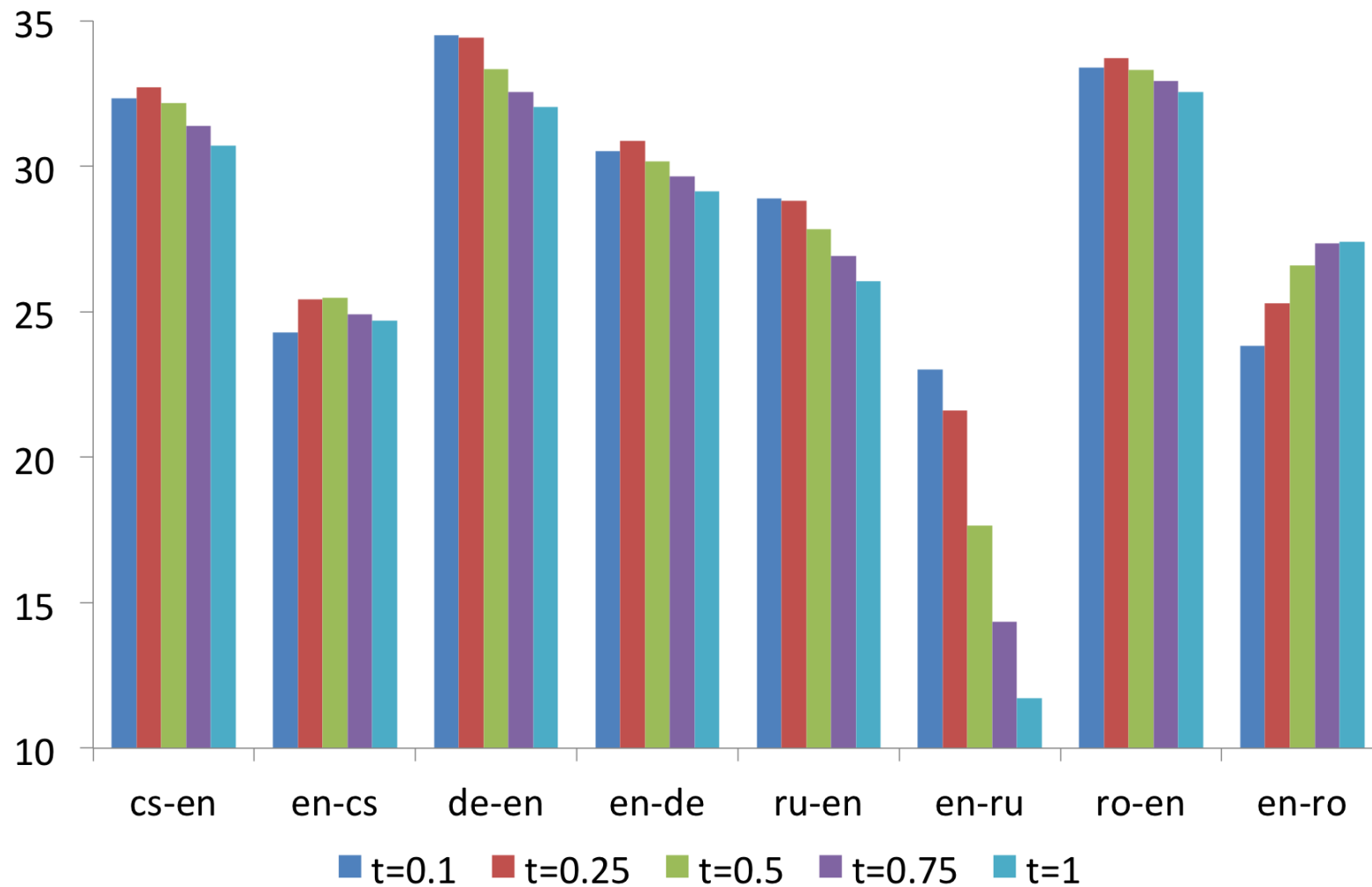


Figure 12

# Domain Adaptation, Low Resource



Graphs here