Continuous Space Reordering Models for Phrase-based MT

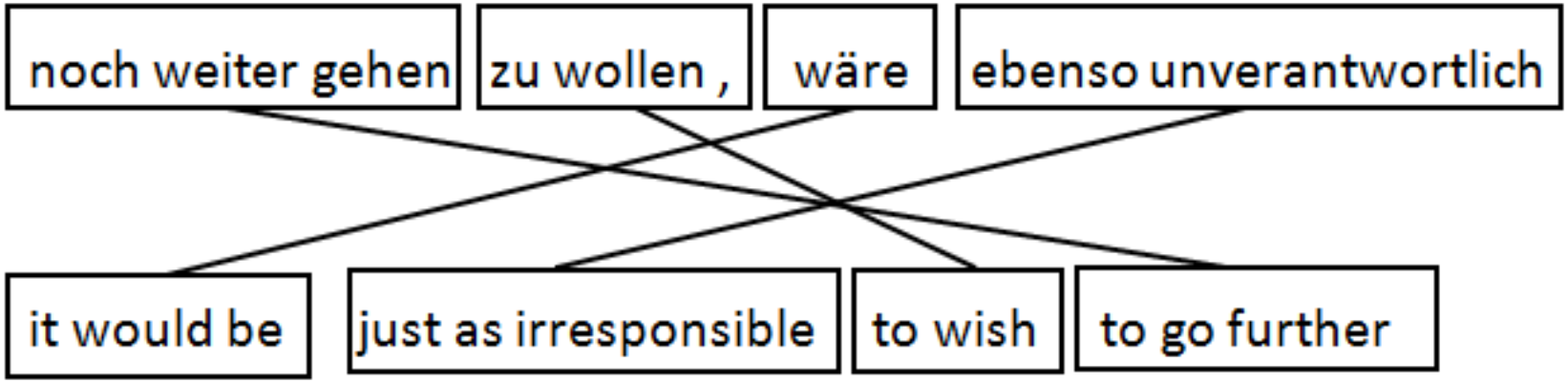
Nadir Durrani and Fahim Dalvi
Qatar Computing Research Institute

In this work

- We neuralize reordering models in Phrase-based using simple Feed-forward NN
  - Lexicalized Reordering Model
  - Operation Sequence Model
  - Modest improvements (+0.5 BLEU points)
- Training Neural MT models with pre-ordered/reordering augmented source
  - Proved harmful

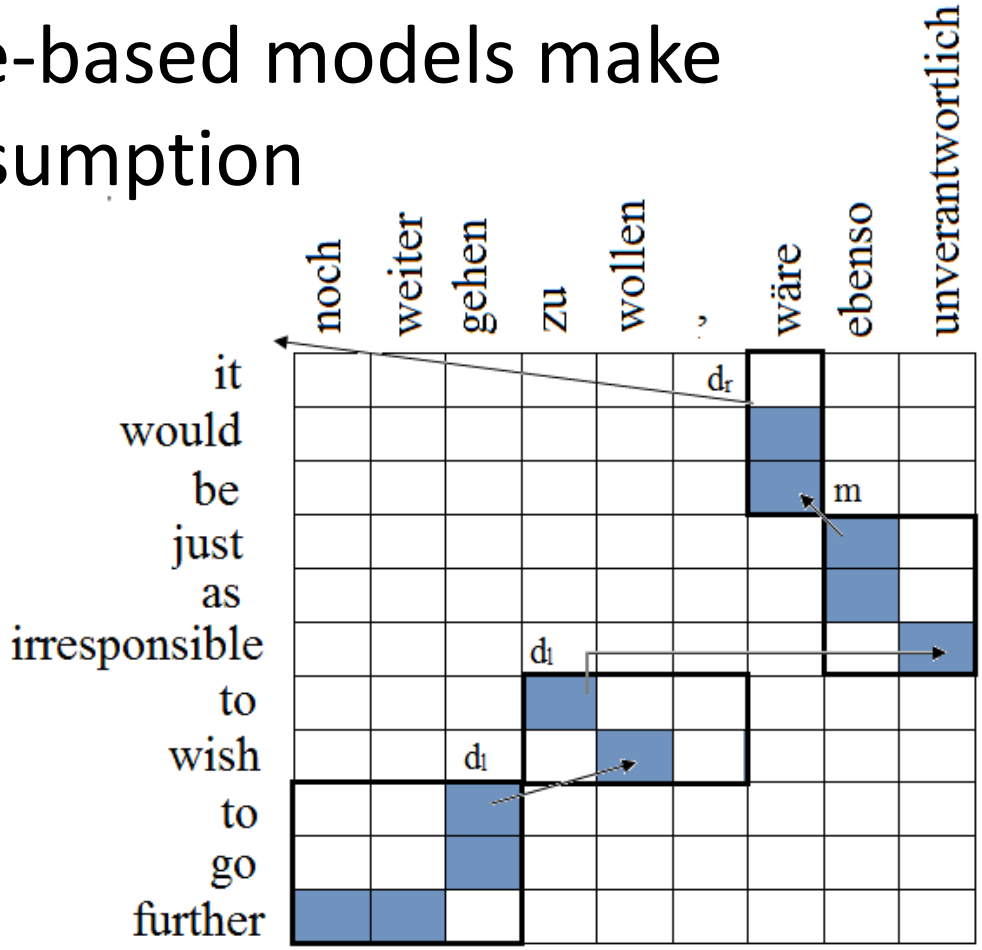
Motivation

- Traditional phrase-based models make independence assumption



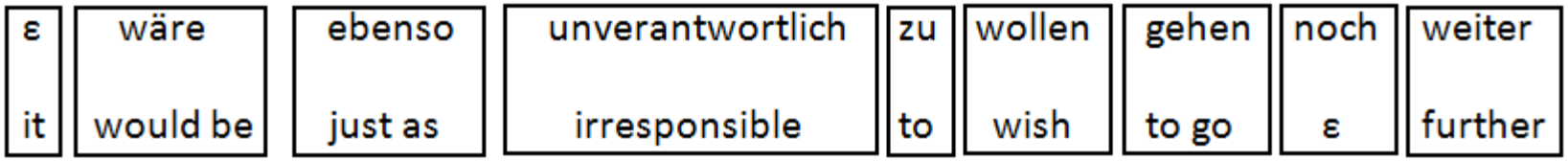
Motivation

- Traditional phrase-based models make independence assumption



Bilingual Markov Models

- Bilingual sequence model addressed this limitation

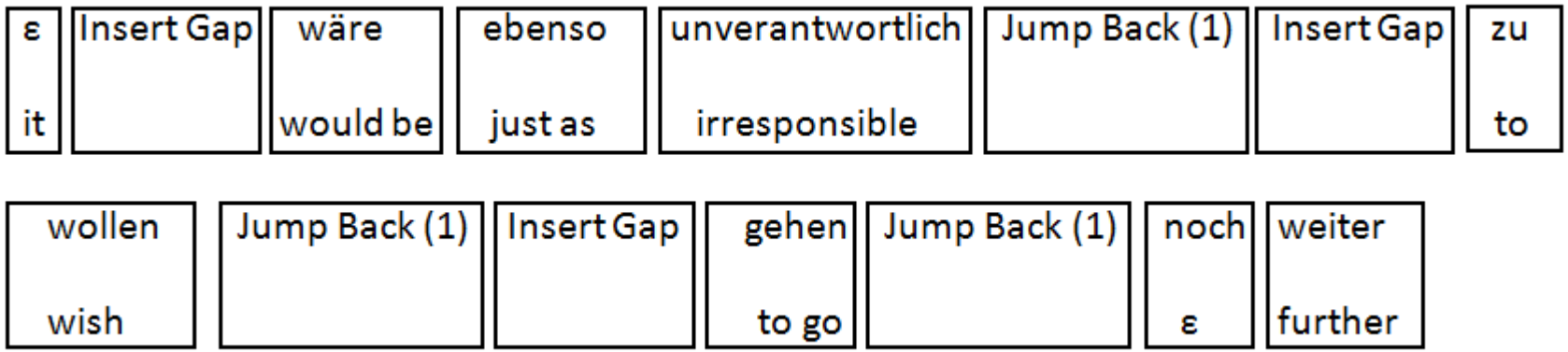


$$p(T,S) \approx \prod_{k=1}^{K} p((t,s)_k | (t,s)_{k-1}, (t,s)_{k-2}, \ldots, (t,s)_{k-n+1})$$

Marino et al., (2006)

Bilingual Markov Models

- Bilingual sequence model with integrated reordering



$$p(T,S) \approx \prod_{j=1}^{J} p(o_j | o_{j-m+1} \cdots o_{j-1})$$

Durrani et al., (2011)

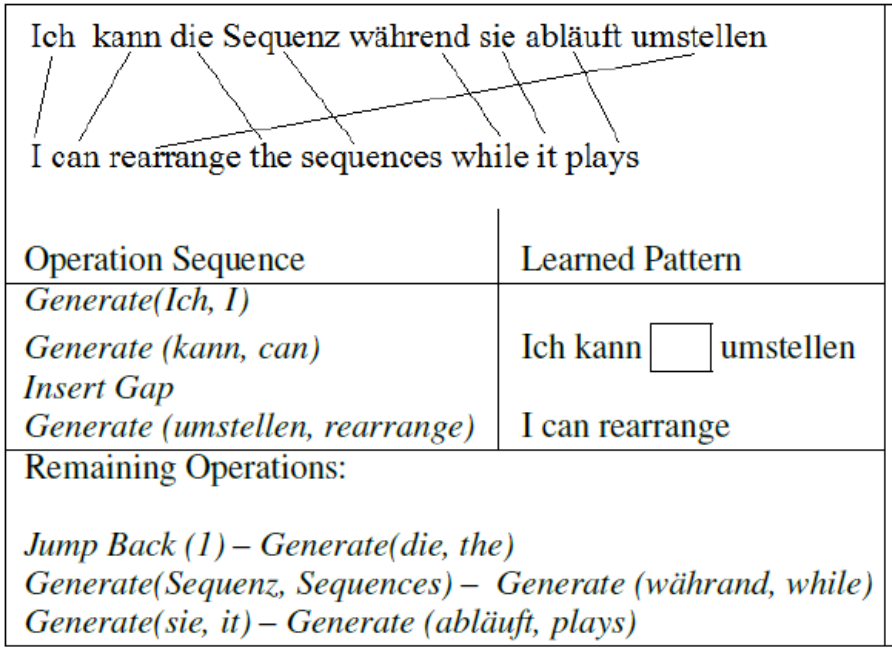
Bilingual Markov Models

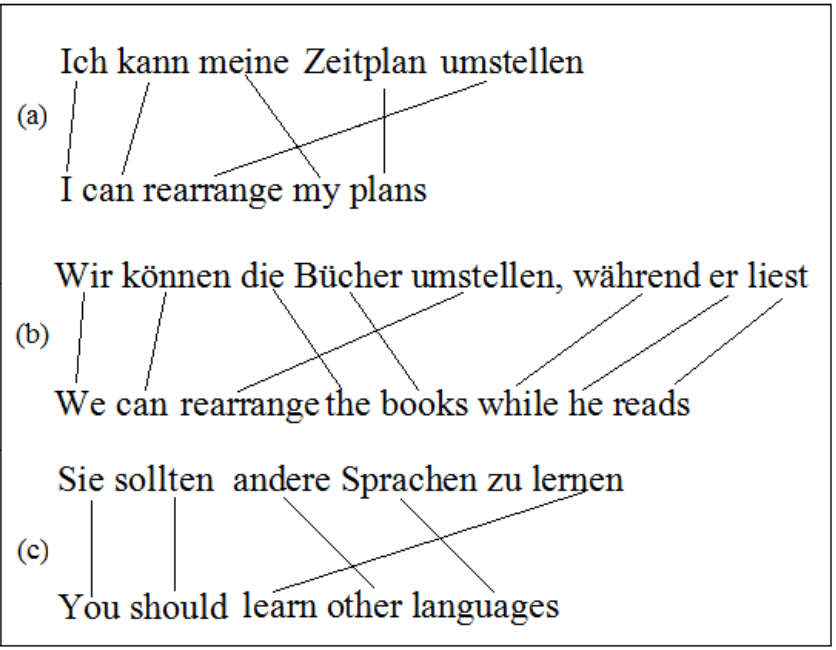
- Based on minimal translation units
  - Overcome phrasal independence assumption
  - No spurious segmentation ambiguity
- Consider source target historical context
- Lexical generation and reordering decisions are tightly coupled

Drawbacks

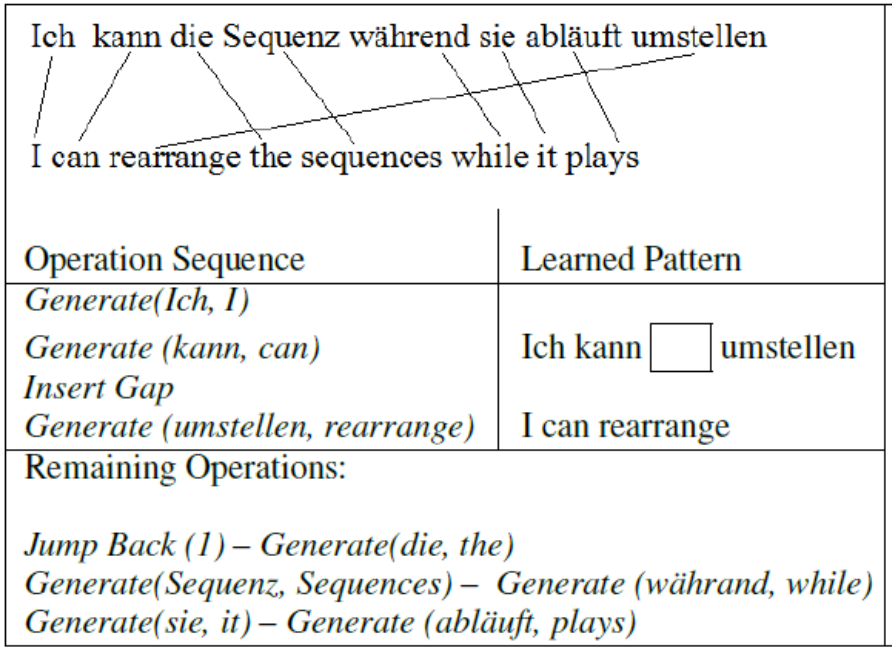
- Due to data sparsity model falls to very short history
  - Joint-ness (combined source target context) contributes to sparsity
  - Rich reordering operations coupled with source-target context
  - Not possible to observe all possible reordering patterns with all possible lexical choices

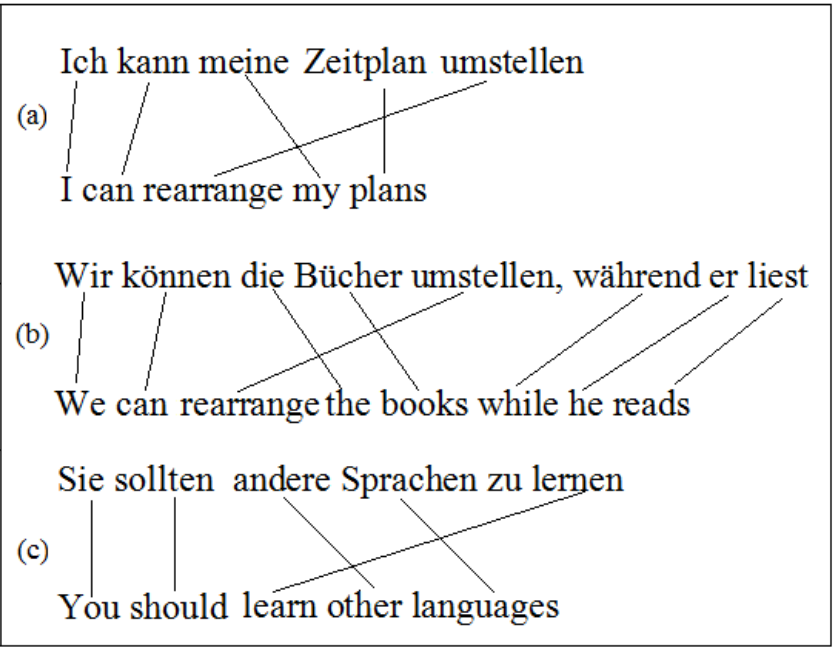
Drawbacks


Ich kann die Sequenz während sie abläuft umstellen
I can rearrange the sequences while it plays
Operation Sequence
Learned Pattern
Generate(Ich, I)
Generate (kann, can)
Insert Gap
Generate (umstellen, rearrange)
Ich kann umstellen
I can rearrange
Remaining Operations:
Jump Back (1) – Generate(die, the)
Generate(Sequenz, Sequences) – Generate (währand, while)
Generate(sie, it) – Generate (abläuft, plays)


(a)
Ich kann meine Zeitplan umstellen
I can rearrange my plans
(b)
Wir können die Bücher umstellen, während er liest
We can rearrange the books while he reads
(c)
Sie sollten andere Sprachen zu lernen
You should learn other languages

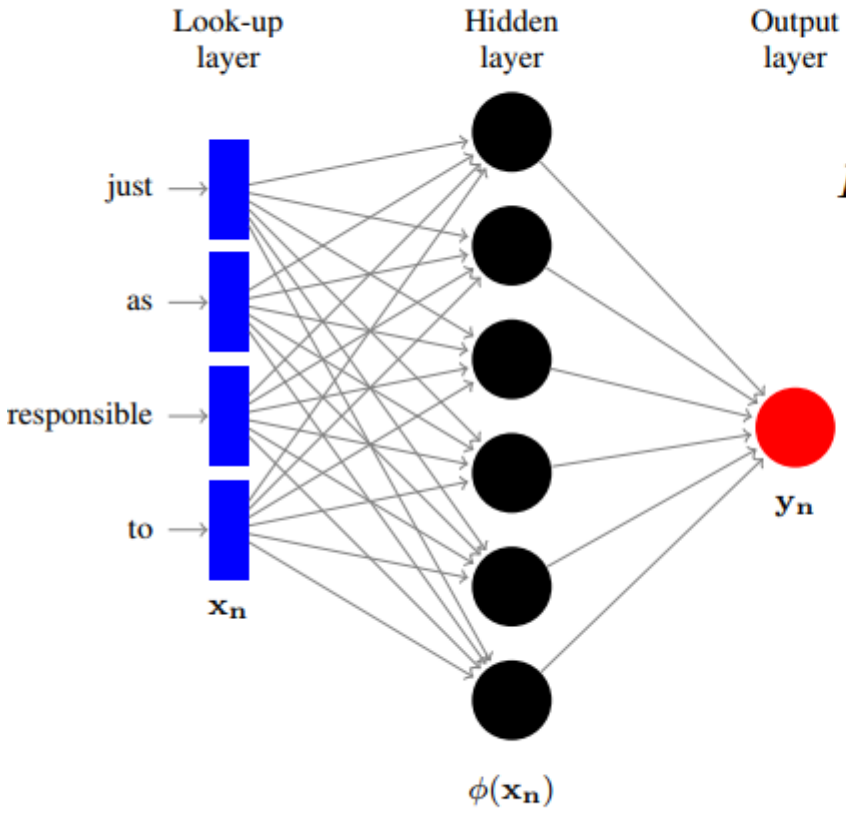
Drawbacks





- Use POS/Morhological Tags/ Word Classes Niehues et al. (2011), Wuebker et al. (2013), Durrani et al. (2014)

Continuous Space Language Model
Bengio et al., (2003)



$$P(y_n = k|\mathbf{x_n}, \theta) = \frac{exp\ (\mathbf{w}_k^T \phi(\mathbf{x_n}))}{\sum_{m=1}^{|V_o|} exp\ (\mathbf{w}_m^T \phi(\mathbf{x_n}))}$$

$$J(\theta) = \sum_{n=1}^{N} \sum_{k=1}^{|V_o|} y_{nk}\ log\ P(y_n = k|\mathbf{x_n}, \theta)$$

Customizations to OSM

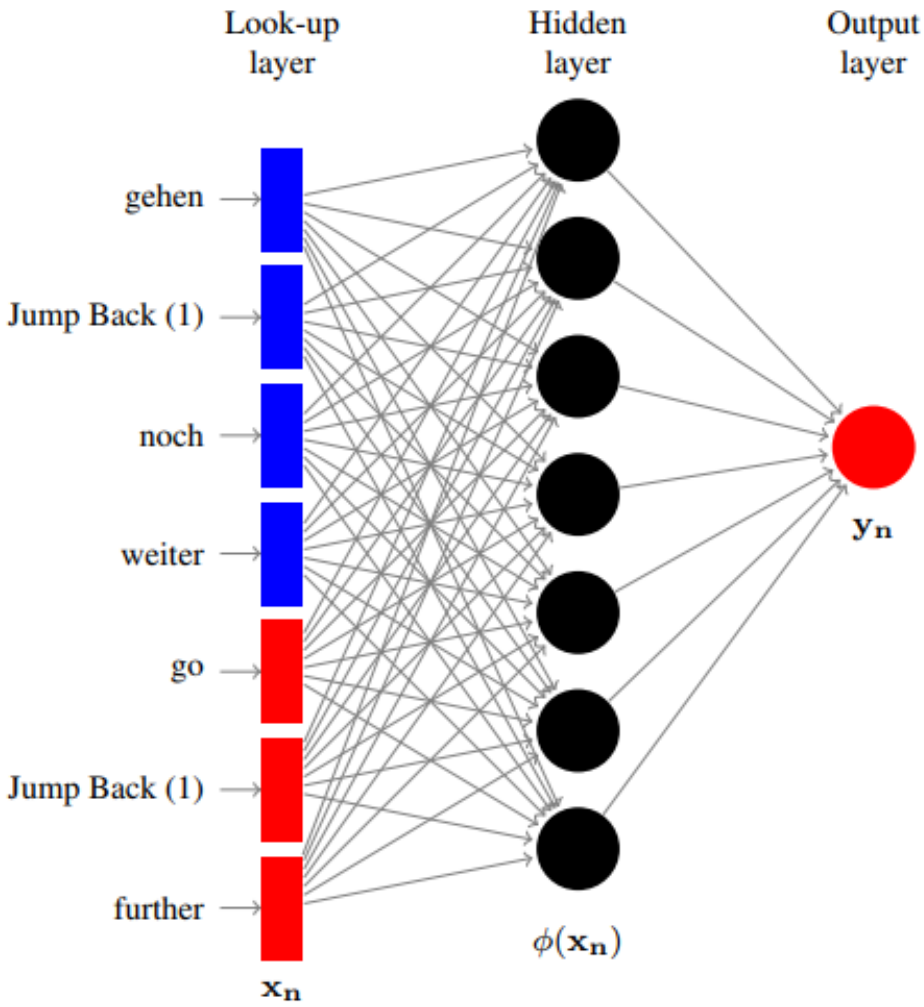
- Because of the joint source target model vocab size becomes quadratic (M x N)
  - Separate out source and target streams and concatenate them to form a hybrid history (Li et., al (2012))
- Reordering operations are appended on both the sides
- Source-side is linearized to be in the target side

Generated Streams for OSM

Operations	Source Stream	Target Stream
Generate Target Only (it)	it	
Insert Gap	Insert Gap	Insert Gap
Generate (wäre, would be)	wäre	would be
Generate (ebenso, just as)	ebenso	just as
Generate (unverantwortlich, irresponsible)	unverantwortlich	irresponsible
Jump Back (1)	Jump Back (1)	Jump Back (1)
Insert Gap	Insert Gap	Insert Gap
Generate (zu, to)	zu	to
Generate (wollen, wish)	wollen	wish
Generate Source Only (,)	,	
Jump Back (1)	Jump Back (1)	Jump Back (1)
Insert Gap	Insert Gap	Insert Gap
Generate (gehen, to go)	gehen	to go
Jump Back (1)	Jump Back (1)	Jump Back (1)
Generate (noch weiter, further)	noch weiter	further

Table 1: Operation Sequences and corresponding streams for Neural OSM

Neural OSM


Look-up layer
Hidden layer
Output layer
gehen
Jump Back (1)
noch
weiter
go
Jump Back (1)
further
$\mathbf{x_n}$
$\phi(\mathbf{x_n})$
$\mathbf{y_n}$

Neural Lexicalized Reordering Model

- Lexicalized Reordering Model
  - Orientation: Monotonic, Swap, Left and Right Discontinuity
  - Task: Predict orientation of a phrase
  - Li et al., (2014)

- Use same machinery
  - Remove monotonic orientation
  - Replace gaps and jumps with left and right discontinuity and swap

Generated Streams for OSM and Lexicalized Reordering Model

Operations	Source Stream	Target Stream	Source Stream	Target Stream
Generate Target Only (it)	it		it	
Insert Gap	Insert Gap	Insert Gap	FD	FD
Generate (wäre, would be)	wäre	would be	wäre	would be
Generate (ebenso, just as)	ebenso	just as	ebenso	just as
Generate (unverantwortlich, irresponsible)	unverantwortlich	irresponsible	unverantwortlich	irresponsible
Jump Back (1)	Jump Back (1)	Jump Back (1)	BD	BD
Insert Gap	Insert Gap	Insert Gap		
Generate (zu, to)	zu	to	zu	to
Generate (wollen, wish)	wollen	wish	wollen	wish
Generate Source Only (,)	,		,	
Jump Back (1)	Jump Back (1)	Jump Back (1)	BD	BD
Insert Gap	Insert Gap	Insert Gap		
Generate (gehen, to go)	gehen	to go	gehen	to go
Jump Back (1)	Jump Back (1)	Jump Back (1)	BD	BD
Generate (noch weiter, further)	noch weiter	further	noch weiter	further

Table 1: Operation Sequences and corresponding streams for Neural OSM and Lexicalized RM training

Evaluation Setup

- Phrase-based Moses
  - Standard settings
  - Including Lexicalized Reordering and OSM Models
- Evaluation Data IWSLT'14
- NN Training (NPLM toolkit)
  - 14-gram (7-gram source + 7-gram target including reordering operations)
  - Vocabulary sizes source: input: 20K, output: 40K
  - Size of Word Vector 150, hidden layer 750
  - Mini-batch: 1000 with 100 noise samples for 25 epochs

Results

- OSM_pos = N-gram OSM model over POS sequences
- OSM_mkcls = N-gram OSM over word clusters
- OSM_neural = Neural OSM
- Lex.reo_neural = Neuralized Lexicalized Reordering with
- Lex_neural = Remove reordering from lexical generation (similar to the Devlin et al,. 2014))

German-English				
System	test11	test12	test13	Avg.
Baseline	35.0	30.3	27.1	30.8
OSM_{pos}	35.3	30.5	27.1	31.0
OSM_{mkcls}	35.1	30.1	26.8	30.7
OSM_{neural}	35.8	31.5	27.0	31.4
$Lex.reo_{neural}$	35.5	31.1	27.2	31.3
Lex_{neural}	35.3	30.8	26.9	31.0
English-German				
Baseline	25.7	21.7	23.4	23.6
OSM_{pos}	25.9	21.9	23.8	23.9
OSM_{mkcls}	25.8	21.8	23.4	23.7
OSM_{neural}	26.1	22.1	24.2	24.1
$Lex.reo_{neural}$	26.1	22.4	23.7	24.1
Lex_{neural}	26.0	22.2	23.7	24.0

Can this help in NMT framework?

- Motivation
  - Reordering is handled solely through *attention model*
    - Cohn et al (2016) aided attention by adding position and structural bias
  - Linearized the source-side can help encoder
  - Reordering operations as encoder states
  - RNN is more powerful than a 14-gram feed-forward NN

Results

- 2-Layered LSTM encoder-decoder

German-English				
System	test11	test12	test13	Avg.
Baseline	33.9	29.2	27.5	30.2
OSM	32.2	27.6	25.6	28.5
Lex.reo	29.2	24.8	22.8	25.6
Lex	30.8	26.6	23.9	27.1

- Du and Way, (2017) also reported that preordering is harmful
  - "Induces noise in terms of word order and hinders the learning process"

Summary

- Neural version of OSM and Lexicalized reordering models
  - Gave +0.5 BLEU improvement over German-English language pairs
  - Better improvements than models trained with POS and mkcls

- Training Neural MT using pre-ordered/reordering augmented data did not help

Thank you !!!