Morphology in MT

March 26th 2015

What is morphology?

- Word formation from smaller parts
- Inflectional
  - eat (V) + -s → eats (V)
- Derivational
  - happy (A) + -ness → happiness (N)
- Compounding
  - dish (N) + washer (N) → dishwasher (N)

What is morphology?

- establish (V)
- disestablish (V)
- disestablishment (N)
- antidisestablishment (N)
- antidisestablishmentary (A)
- antidisestablishmentarian (N)
- antidisestablishmentarianism (N)

What is morphology?

Unabhängigkeitserklärung

我们

reunification

библиотеку

입혔습니까

kitapçığa

শব্দাবলী

聞かせられたら

להבדיל

inquiriendo

étudiiez

tusaatsiarunnanngittualuujunga

Problems

- Alignment
- Phrase scoring
- Input OOVs
- Novel form generation
- Language Modeling
- Evaluation

Alignment

green colorless ideas sleep
bezbarvé zelené myšlenky spí

I like green pears
mám rád zelené hrušky

I sat under a green tree
seděl jsem pod zeleným stromem

Alignment

green colorless ideas sleep
bezbarvé zelené myšlenky spí

I like green pears
mám rád zelené hrušky

I sat under a green tree
seděl jsem pod zeleným stromem

Phrase Scoring

en	cs	p(cs l en)
cat	kočka	0.5629
cat	kočku	0.1769
cat	kočce	0.0002
cat	kočky	0.00004

en	cs	p(cs l en)
cat	kokour	0.112
cat	kokoura	0.074
cat	kokourovi	0.017
cat	kokoura	0.0051

Phrase Scoring

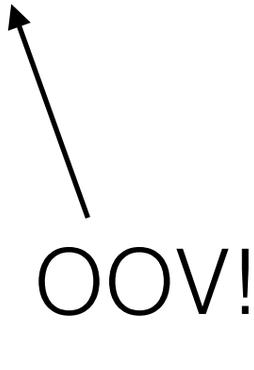
en	cs	p(cs l en)
cat	kočka	0.7805
cat	kokour	0.2194

en	cs	p(cs l en)
cat	NOM	0.7117
cat	ACC	0.2646
cat	DAT	0.018
cat	GEN	0.0054

Input OOVs

- La mejor aplicación sería la que **erradicase** el hambre del mundo.

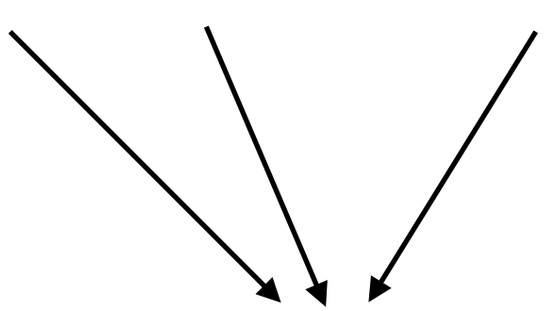
f	e	p(elf)
la	the	0.9173
mejor	best	0.6330
aplicación	application	0.8211
ser	to be	0.1182
sería	would be	0.3442
la que	to	0.0596
erradica	eradicates	0.9754
erradicó	eradicated	0.9303
erradican	eradicate	0.9481
erradico	erradicate	0.8731
erradicando	eradicating	0.9713
el hambre	hunger	0.5385
del mundo	world	0.2006



- The best application would be to erradicase world hunger. 😒

Novel Form Generation

She had **attempted** to cross the road on her bike.



Она **пыталась** пересечь пути на её велосипед.

Language Modeling

- Je **porte un parapluie** dans mon sac .
- À Seattle , on doit **porter un parapluie** tous les jours .

Language Modeling

- Je **porte un parapluie** dans mon sac .
- À Seattle , on doit **porter un parapluie** tous les jours .
- PRN **VB DT NN** PREP PRP$ NN PUNC
- PREP NNP PUNC PRP VB **VB DT NN** PDT DT NNS .

Evaluation

	Sentence	BLEU +1
Input	The earnings on its 10 - year bonds are 28.45 % .	-
Reference	Výnos na jejích 10 - letých dluhopisech je na 28,45 % .	100.00
System 1	Příjmy na své desetileté dluhopisy jsou 28,45 % .	22.61
System 2	Příjmy na jeho 10 - letých poutech jsou 28,45 % .	32.04

Evaluation

	Sentence	BLEU +1
Input	The earnings on its 10 - year bonds are 28.45 % .	-
Reference	Výnos na jejích 10 - letých dluhopisech je na 28,45 % .	100.00
System 1	Příjmy na své desetileté dluhopisy jsou 28,45 % .	22.61
System 2	Příjmy na jeho 10 - letých poutech jsou 28,45 % .	32.04
Another Human	Zisk z jejích 10 - letých dluhopisů je 28,45 % .	32.04

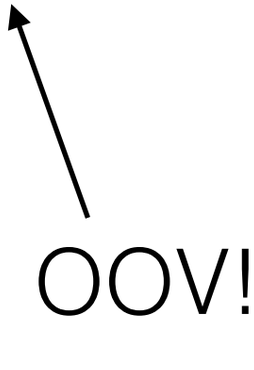
Overview

- Morphology on the source side
  - Stemming
  - Lattices
- Morphology on the target side
  - Source enrichment
  - Factored models
  - Synthetic phrases
  - Other formalisms?

Stemming

- La mejor aplicación sería la que **erradicase** el hambre del mundo.

f	e	p(elf)
la	the	0.9173
mejor	best	0.6330
aplicación	application	0.8211
ser	to be	0.1182
sería	would be	0.3442
la que	to	0.0596
erradica	eradicates	0.97540
erradicó	eradicated	0.9303
erradican	eradicate	0.9481
erradico	erradicate	0.8731
erradicando	eradicating	0.9713
el hambre	hunger	0.5385
del mundo	world	0.2006



- The best application would be to erradicase world hunger. 😒

Stemming

- La mejor aplic ser la que **erradic** el hambr del mundo.

f	e	p(elf)
la	the	0.9173
mejor	best	0.6330
aplic	application	0.8211
ser	to be	0.0807
ser	would be	0.0338
la que	to	0.0596
erradic	eradicates	0.0633
erradic	eradicated	0.2173
erradic	eradicate	0.3880
erradic	eradicating	0.1503
el hambr	hunger	0.5385
del mundo	world	0.2006

- The best application to be to eradicate world hunger. 😐

Stemming

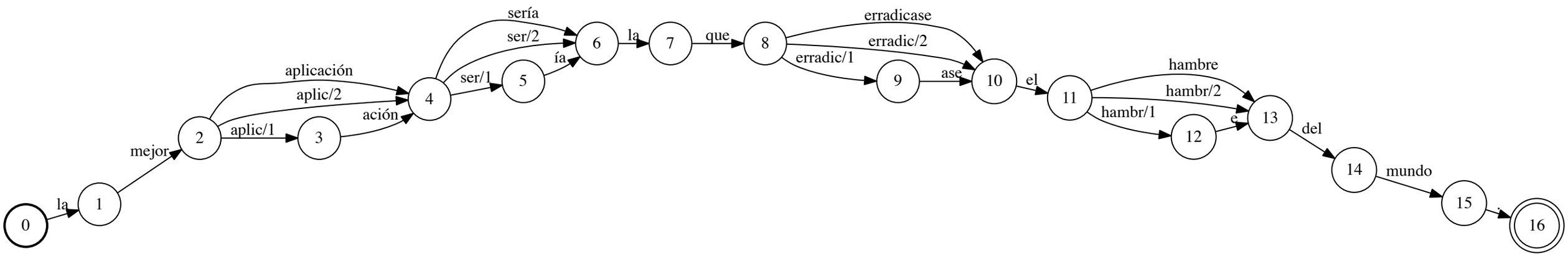
- La mejor aplic ación ser ía la que **erradic ase** el hambr e del mundo.

f	e	p(elf)
la	the	0.9173
mejor	best	0.6330
aplic ación	application	0.8211
ser	to be	0.1182
ser ía	would be	0.3442
la que	to	0.0596
erradic a	eradicates	0.97540
erradic ó	eradicated	0.9303
erradic an	eradicate	0.9481
erradic o	erradicate	0.8731
erradic ando	eradicating	0.9713
el hambr e	hunger	0.5385
del mundo	world	0.2006

f	e	p(elf)
erradic	erradicate	0.2571
erradic	erradicating	0.1253
ase	have been	0.1334

- The best application would be to have been eradicating world hunger. 😐

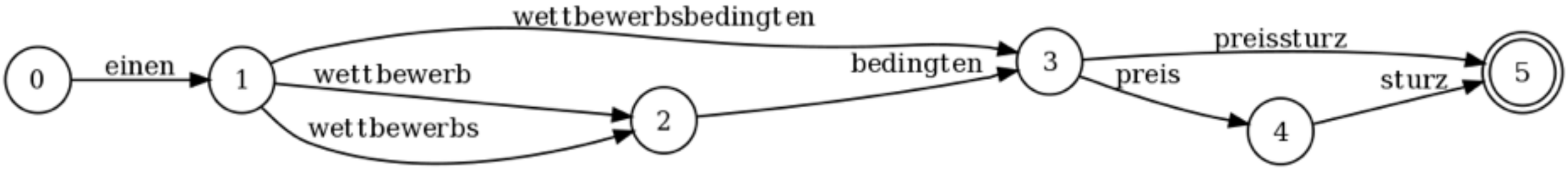
Input Lattices



- The best application would be to eradicate world hunger. 😀

Dyer et. al (2008)

Input Lattices



"a competition-induced price fall"

Dyer et. al (2008)

Take Aways

- If you have source side morphology:
  - Stem, lowercase, and compound split your data when doing alignment
  - Extract phrases normally
  - Use input lattices during tuning and decoding

Target Side Morphology

I want to eat a sandwich .

Chci jíst sendvič .

Source Enrichment

Avramidis and Koehn (2008)

Factored Models

Surface

neue

häuser

werden

gebaut

Koehn and Hoang (2007)

Factored Models

Surface	Lemma	POS	Morph.
neue	neu	JJ	+pl +fem
häuser	häus	NN	+pl
werden	werden	VB	+3+pl +pres
gebaut	bauen	VBN	+past +part

Koehn and Hoang (2007)

Factored Models

Surface	Lemma	POS	Morph.
neue	neu	JJ	+pl +fem
häuser	häus	NN	+pl
werden	werden	VB	+3+pl +pres
gebaut	bauen	VBN	+past +part

Surface	Lemma	POS	Morph.

Koehn and Hoang (2007)

Factored Models

Surface	Lemma	POS	Morph.
neue	**neu**	JJ	+pl +fem
häuser	**häus**	NN	+pl
werden	**werden**	VB	+3+pl +pres
gebaut	**bauen**	VBN	+past +part

Surface	Lemma	POS	Morph.

Koehn and Hoang (2007)

Factored Models

Surface	Lemma	POS	Morph.
neue	**neu**	JJ	+pl
häuser	**häus**	NN	
werden	**werden**	VB	
gebaut	**bauen**	VBN	+past +part

Surface	Lemma	POS	Morph.

DE	EN	p(EN |
häus	house	0.76
häus	home	0.15
haüs	buildin	0.06
haüs	shell	0.02

Koehn and Hoang (2007)

Factored Models

Surface	Lemma	POS	Morph.
neue	**neu**	JJ	+pl +fem
häuser	**häus**	NN	+pl
werden	**werden**	VB	+3+pl +pres
gebaut	**bauen**	VBN	+past +part

Surface	Lemma	POS	Morph.
new			
house			
be			
build			

Koehn and Hoang (2007)

Factored Models

Surface	Lemma	POS	Morph.
neue	neu	**JJ**	**+pl +fem**
häuser	häus	**NN**	**+pl**
werden	werden	**VB**	**+3+pl +pres**
gebaut	bauen	**VBN**	**+past +part**

Surface	Lemma	POS	Morph.
new			
house			
be			
build			

Koehn and Hoang (2007)

Factored Models

Surface	Lemma	DE	EN	p(EN | DE)	POS	Morph.
neue	neu	VB+3p+pl+pres	VB+3p+pl+pres	0.81		
		VB+3p+pl+pres	VB+3p+sg+pres	0.10		
		VB+3p+pl+pres	PRN+3p+pl	0.04		
		VB+3p+pl+pres	NN+pl	0.03		
häuser	häus				**NN**	**+pl**
werden	werden				**VB**	**+3+pl +pres**
gebaut	bauen				**VBN**	**+past +part**

house

be

build

Koehn and Hoang (2007)

Factored Models

Surface	Lemma	POS	Morph.
neue	neu	**JJ**	**+pl +fem**
häuser	häus	**NN**	**+pl**
werden	werden	**VB**	**+3+pl +pres**
gebaut	bauen	**VBN**	**+past +part**

Surface	Lemma	POS	Morph.
new		**JJ**	
house		**NN**	**+pl**
be		**VB**	**+3+pl +pres**
build		**VBN**	**+past +part**

Koehn and Hoang (2007)

Factored Models

Surface	Lemma	POS	Morph.
new		JJ	
house		NN	+pl
be		VB	+3+pl +pres
build		VBN	+past +part

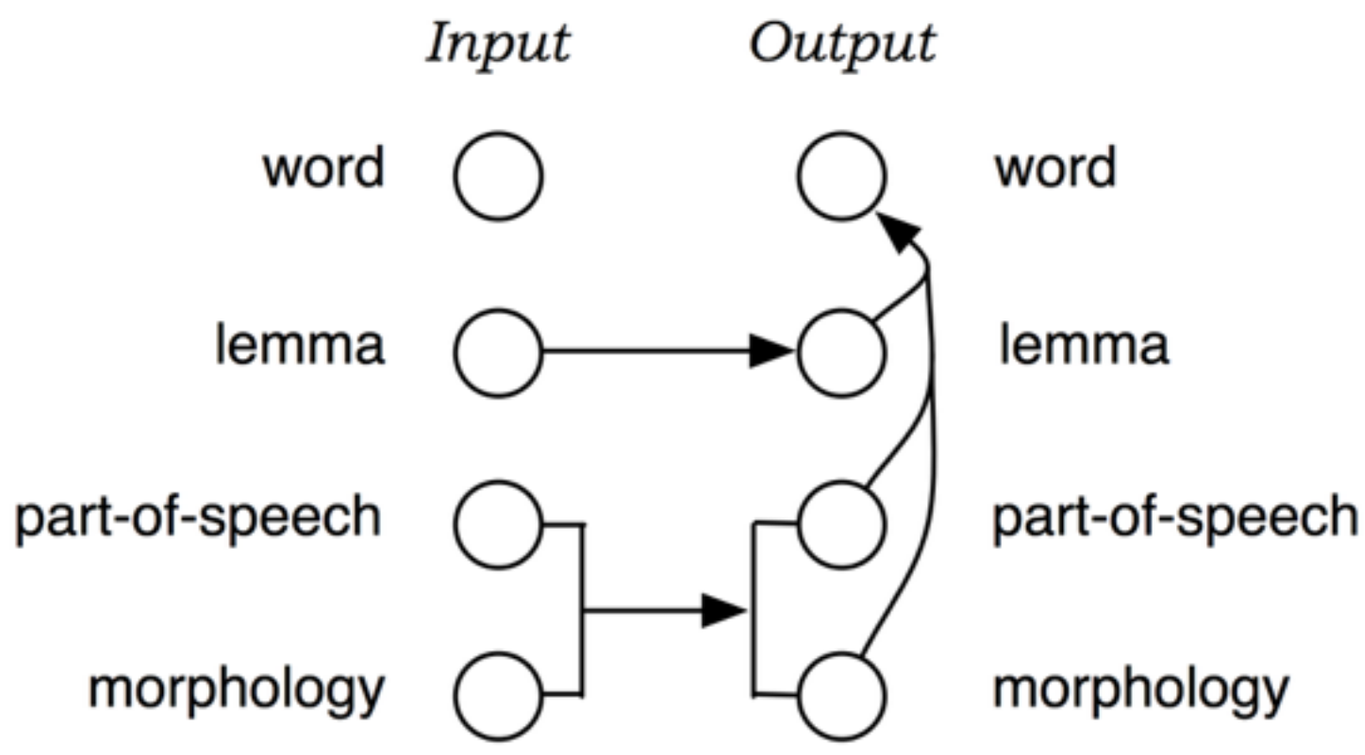
Koehn and Hoang (2007)

Factored Models

Surface	Lemma	POS	Morph.
new	new	JJ	
houses	house	NN	+pl
are	be	VB	+3+pl +pres
built	build	VBN	+past +part

Koehn and Hoang (2007)

Factored Models



Koehn and Hoang (2007)

Factored Models

Pros	Cons
Much more human-like	Huge search space
Can generate novel forms	Word forms generated independently
Factored language models	Changes whole MT Pipeline

Koehn and Hoang (2007)

Factored Models

Pros	Cons
Much more human-like	Huge search space
Can generate novel forms	Word forms generated independently
Factored language models	Changes whole MT Pipeline

Koehn and Hoang (2007)

POS Language Models

- Convert corpus to POS tags instead of surface forms

- Build large (7~8) n-gram models

The president announced his new plan yesterday .

The council approved the sanctions on Iran .

Polls in the UK show the LDP up 2 % over last year .

POS Language Models

- Convert corpus to POS tags instead of surface forms
- Build large (7~8) n-gram models

DT NN VBD PRP$ JJ NN ADV PUNC
DT NN VBD DT NN IN NNP PUNC
NNS IN DT NNP VB DT NNP NUM PUNC IN JJ NN PUNC

Brown Cluster LMs

- Automatically induce word classes
- Can capture more nuances of the language

The president announced his new plan yesterday .

The council approved the sanctions on Iran .

Polls in the UK show the LDP up 2 % over last year .

Brown Cluster LMs

- Automatically induce word classes
- Can capture more nuances of the language
- Example

The president announced his new plan yesterday .

The council approved the sanctions on Iran .

Polls in the UK show the LDP up 2 % over last year .

Brown Cluster LMs

- Automatically induce word classes
- Can capture more nuances of the language
- Example

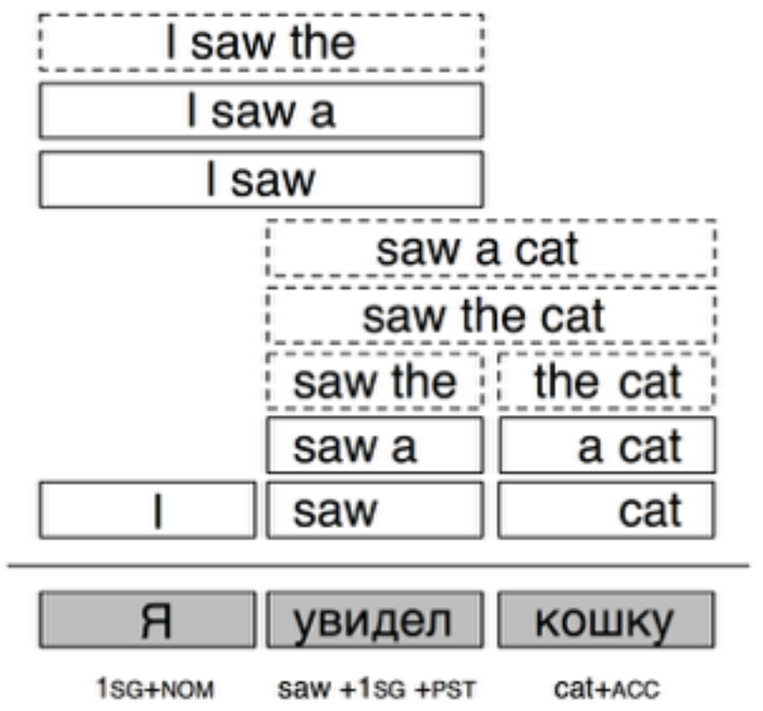
10010 110110010 0100111110010
10011111 1010011100
110111111010 01011011111
000000

10010 1101101011110
0100111100111 10010
110100111111 001110110
111100011 000000

110001110 0011100 10010
11111101011 01000001111 10010
1111110100 010110000 111101011
1111101110 001110111111110
10111010 1111100101 000000

Synthetic Phrases

- Dynamically add phrases to the translation table
- Can condition on source sentence, phrase table, and more!
- Originally used to insert determiners in RU →EN translation



Tsvetkov et al. (2013)

Synthetic Phrases

σ:пытаться_V + μ:mis-sfm-e

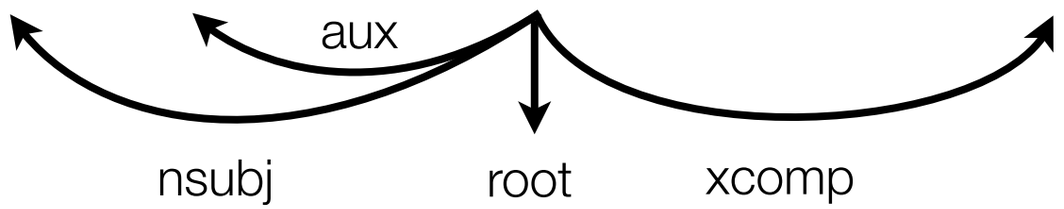
она **пыталась** пересечь пути на ее велосипед

-1 +1

she had attempted to cross the road on her bike

C50 C473 C28 C8 C275 C37 C43 C82 C94 C331

PRP VBD VBN TO VB DT NN IN PRP$ NN



Chahuneau et al. (2013)

Synthetic Phrases

- Generate compound words in the target language
- Character-level MT system

Matthews et al. (2015*)

Synthetic Phrases

- Generate compound words in the target language
- Character-level MT system

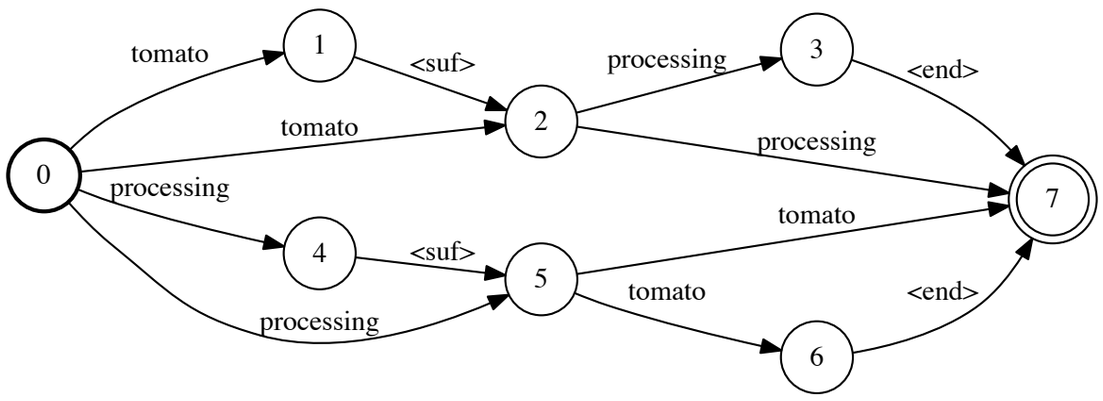
EN	DE	Score
tomato	t o m a t e	-2.58
processing	v e r a r b e i t	-0.75
processing	b e h a n d l u n g	-2.74
processing	v e r e d e l u n g	-4.94

EN	DE	Score
<suf>	n	-3.71
<suf>	s	-2.53
<end>	u n g	-5.73
<end>	e n d e	-9.86

Matthews et al. (2015*)

Synthetic Phrases

- Generate compound words in the target language
- Character-level MT system



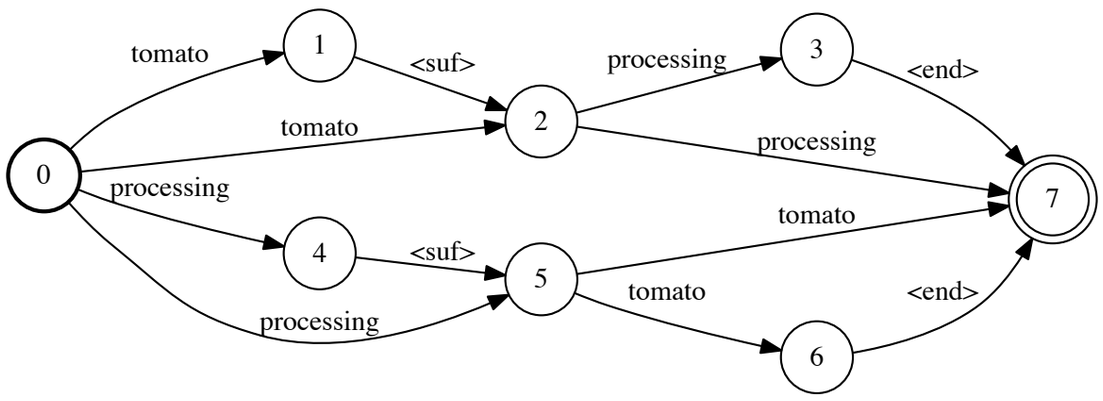
EN	DE	Score
tomato	t o m a t e	-2.58
processing	v e r a r b e i t	-0.75
processing	b e h a n d l u n g	-2.74
processing	v e r e d e l u n g	-4.94

EN	DE	Score
<suf>	n	-3.71
<suf>	s	-2.53
<end>	u n g	-5.73
<end>	e n d e	-9.86

Matthews et al. (2015*)

Synthetic Phrases

- Generate compound words in the target language
- Character-level MT system



tomatenverarbeitung

EN	DE	Score
tomato	t o m a t e	-2.58
processing	v e r a r b e i t	-0.75
processing	b e h a n d l u n g	-2.74
processing	v e r e d e l u n g	-4.94

EN	DE	Score
<suf>	n	-3.71
<suf>	s	-2.53
<end>	u n g	-5.73
<end>	e n d e	-9.86

Matthews et al. (2015*)

More Ideas

- Use information to synthetically add/modify feature values
- Add synthetic phrases for discourse-level information
- Add synthetic grammar rules in addition to phrase pairs
- Many, many more!

Take Aways

- Use Brown Cluster LMs (c=600, o=7)
  - (whether you have target morphology or not!)
- Synthetic phrases can solve a wide range of target-side generation problems
- Check out morphogen