modeling morphological uncertainty for machine translation

Waleed Ammar, Victor Chahuneau, Chris Dyer, Noah A. Smith (CMU)

Jason Baldridge, Kyle Jerro, Jason Mielens (UT)

three observations

rich morphology is typical

morphology is a challenge
(esp. into rich)

linguistic knowledge can help

translation puzzle

English	**Kinyarwanda**	**English**	**Kinyarwanda**
big man	umugabo munini	small man	umugabo mutoya
big women	? ?	small women	abagore batoya
big boat	ubwato bunini	small boat	ubwato butoya
big word	ijambo rinini	small word	ijambo ritoya
big thing	ikintu kinini	small thing	? ?
big things	ibintu binini	small things	ibintu bitoya
big cow	inka nini	small cow	inka toya

linguists

English	Kinyarwanda	English	Kinyarwanda
big man (class 1)	umugabo munini	small man	umugabo mutoya
big women (class 2)	? ?	small women	abagore batoya
big boat (class 14)	ubwato bunini	small boat	ubwato butoya
big word (class 5)	ijambo rinini	small word	ijambo ritoya
big thing (class 7)	ikintu kinini	small thing	? ?
big things (class 8)	ibintu binini	small things	ibintu bitoya
big cow (class 9)	inka nini	small cow	inka toya
...	...	...	...

linguists

English	Kinyarwanda	English	Kinyarwanda
big man (class 1)	umugabo munini	small man	umugabo mutoya
big women (class 2)	abagore class2(nini)	small women	abagore batoya
big boat (class 14)	ubwato bunini	small boat	ubwato butoya
big word (class 5)	ijambo rinini	small word	ijambo ritoya
big thing (class 7)	ikintu kinini	small thing	ikintu class7(toya)
big things (class 8)	ibintu binini	small things	ibintu bitoya
big cow (class 9)	inka nini	small cow	inka toya
...	...	...	...

linguists

English	Kinyarwanda	English	Kinyarwanda
big man (class 1)	umugabo munini	small man	umugabo mutoya
big women (class 2)	abagore banini ✓	small women	abagore batoya
big boat (class 14)	ubwato bunini	small boat	ubwato butoya
big word (class 5)	ijambo rinini	small word	ijambo ritoya
big thing (class 7)	ikintu kinini	small thing	ikintu kitoya ✓
big things (class 8)	ibintu binini	small things	ibintu bitoya
big cow (class 9)	inka nini	small cow	inka toya
...	...	...	...

mainstream translation models

English	**Kinyarwanda**	**English**	**Kinyarwanda**
big man	umugabo munini	small man	umugabo mutoya
big women	? ?	small women	abagore batoya
big boat	ubwato bunini	small boat	ubwato butoya
big word	ijambo rinini	small word	ijambo ritoya
big thing	ikintu kinini	small thing	? ?
big things	ibintu binini	small things	ibintu bitoya
big cow	inka nini	small cow	inka toya
...	...	...	...

mainstream word alignment methods

English	**Kinyarwanda**	**English**	**Kinyarwanda**
big man	umugabo munini	small man	umugabo mutoya
big women	? ?	small women	abagore batoya
big boat	ubwato bunini	small boat	ubwato butoya
big word	ijambo rinini	small word	ijambo ritoya
big thing	ikintu kinini	small thing	? ?
big things	ibintu binini	small things	ibintu bitoya
big cow	inka nini	small cow	inka toya
...	...	...	...

mainstream word alignment methods

English	**Kinyarwanda**	**English**	**Kinyarwanda**
big man	umugabo munini	small man	umugabo mutoya
big women	? ?	small women	abagore batoya
big thing	ikintu kinini	small thing	? ?
big things	ibintu binini	small things	ibintu bitoya
big cow	inka nini	small cow	inka toya
...	...	...	...

Problem #1: word **alignment** errors for rare inflections.

decoding

English	Kinyarwanda	English	Kinyarwanda
big man	umugabo munini	small man	umugabo mutoya
big women	tr(women) tr(big)	small women	abagore batoya
big boat	ubwato bunini	small boat	ubwato butoya
big word	ijambo rinini	small word	ijambo ritoya
big thing	ikintu kinini	small thing	tr(thing) tr(small)
big things	ibintu binini	small things	ibintu bitoya
big cow	inka nini	small cow	inka toya
...	...	...	...

decoding

English	Kinyarwanda	English	Kinyarwanda
big man	umugabo munini	small man	umugabo mutoya
big women	abagore munini	small women	abagore batoya
big boat	ubwato banini	small boat	ubwato butoya
big word	ijambo rinini	small word	ijambo ritoya
big thing	ikintu kinini	small thing	ikintu mutoya
big things	ibintu binini	small things	ibintu kitoya
big cow	inka nini	small cow	inka toya
...	...	...	...

decoding

English	Kinyarwanda	English	Kinyarwanda
big man	umugab banini	small man	umugabo mutoya
big women	abagore munini	small women	abagore batoya
big thing	ikintu kinini	small thing	ikintu mutoya
big things	ibintu binini	small things	ibintu kitoya
big cow	inka nini	small cow	inka toya
...	...	...	...

Problem #2: inability to **generate** unseen inflections.

decoding

English | **Kinya** | **Kinyarwanda**

real problems

Problem #1: word **alignment** errors for rare inflections.

Problem #2: inability to **generate** unseen inflections.

Problem #3: n-gram LM fails to **predict** rare inflections.

Goal: use linguistic knowledge to overcome #1, #2 and #3.

... | ... | ... | ...

linguistic resources

- morphological analyzer(s) for the rich language

road map

Problem #1: word alignment errors for rare inflections.

Problem #2: inability to generate unseen inflections.

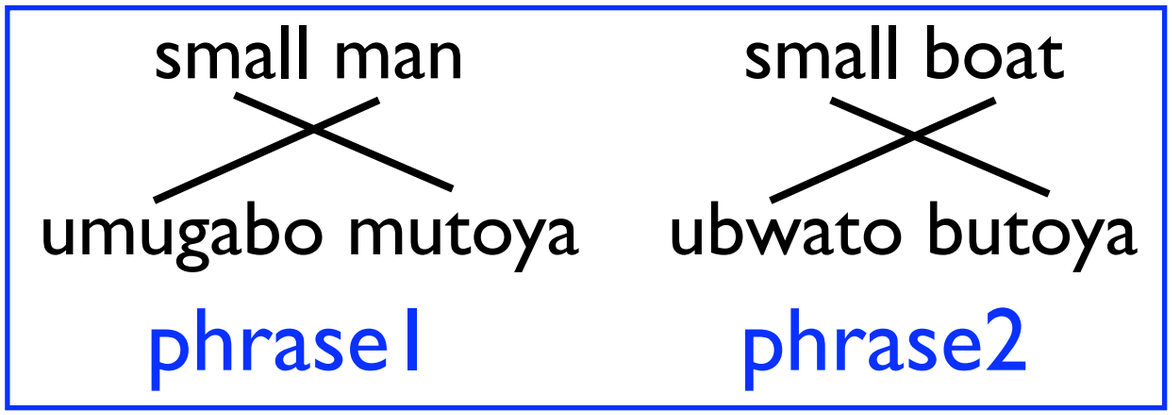
Problem #3: n-gram LM fails to predict rare inflections.

Problem #1: word alignment errors for rare inflections.

unsupervised word alignment with morphological features

more general features

$$\frac{\exp \boldsymbol{\theta}^\top \mathbf{H}(\mathbf{t}, \mathbf{a}, \mathbf{s}, n)}{Z_{\boldsymbol{\theta}}(\mathbf{s}, n)}$$



- HMM and IBM models
  - "small-mutoya" phrase1
  - "small-butoya" phrase2
- log-linear models
  - "small-mutoya" phrase1
  - "small-butoya" phrase2
  - "small-□toya" phrase1,2

Problem #1: word alignment errors for rare inflections.

unsupervised word alignment with morphological features

need to compute feature expectations

too slow, even with small amounts of data

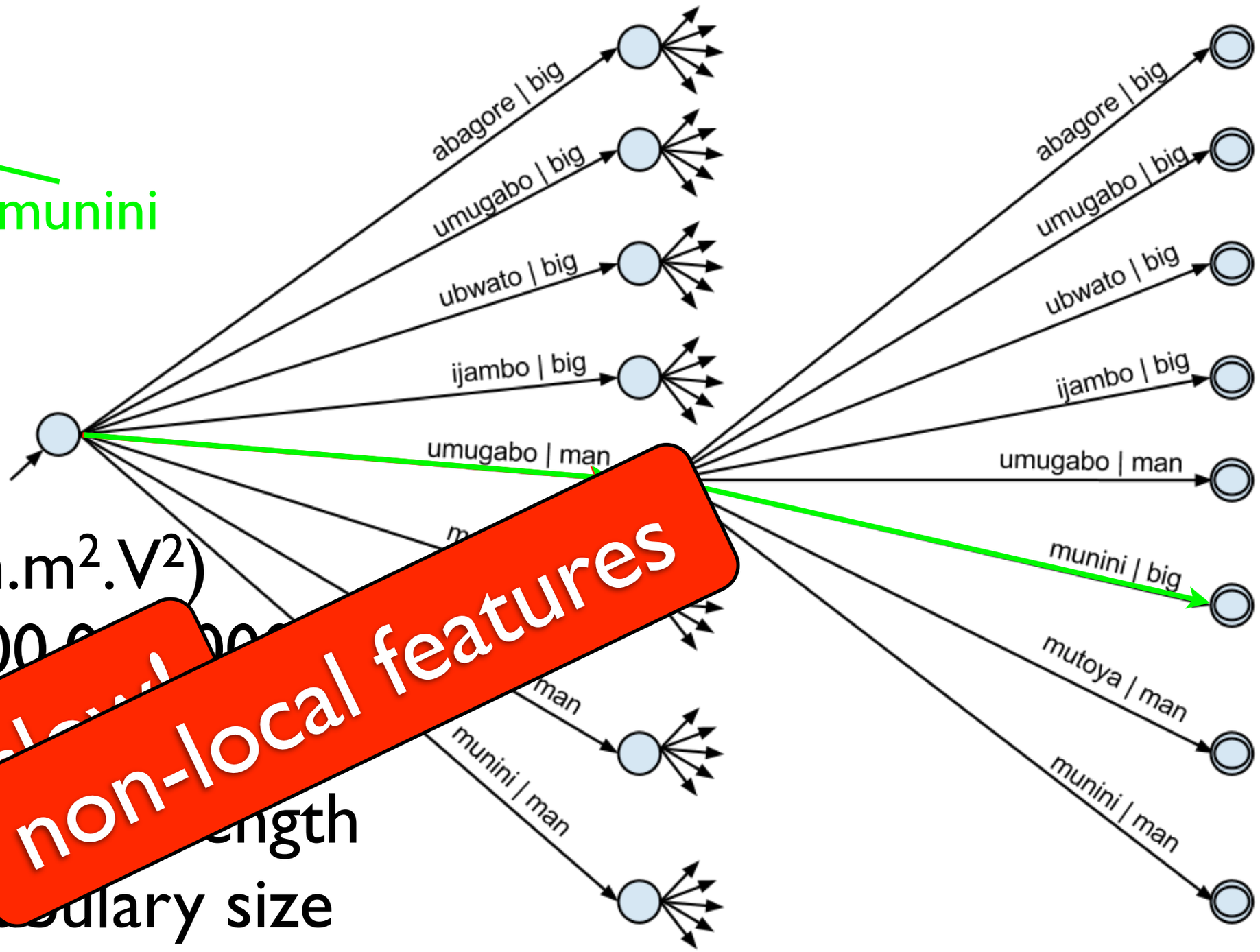
- 2011:
  - finite state transducers
  - discriminative lexicon
- 2012:
  - approximate inference
  - non-local features
  - OpenFST

Problem #1: word alignment errors for rare inflections.

alignment FST

Eng: big man

Kin: umugabo munini



arcs = O(n.m².V²)

~ 1,000 [illegible]

n: source [illegible]

m: ta[illegible]ength

V: target vocabulary size

too slow!

non-local features

Problem #1: word alignment errors for rare inflections.

alignment FST

Eng: big man

proposal distribution

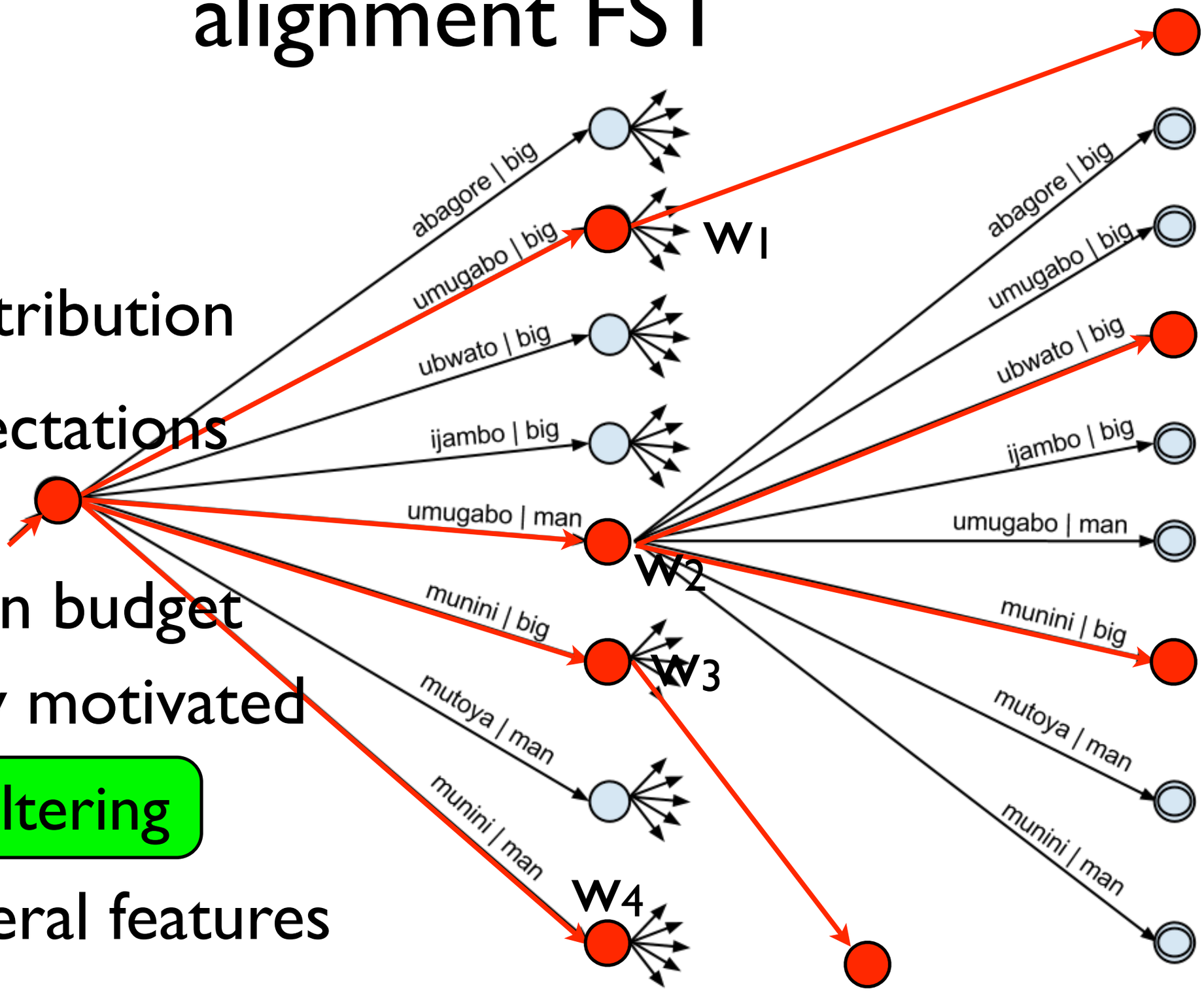
feature expectations

computation budget

theoretically motivated

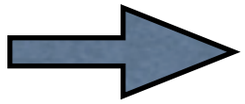
particle filtering

enables general features



road map

Problem #1: word alignment errors for rare inflections.



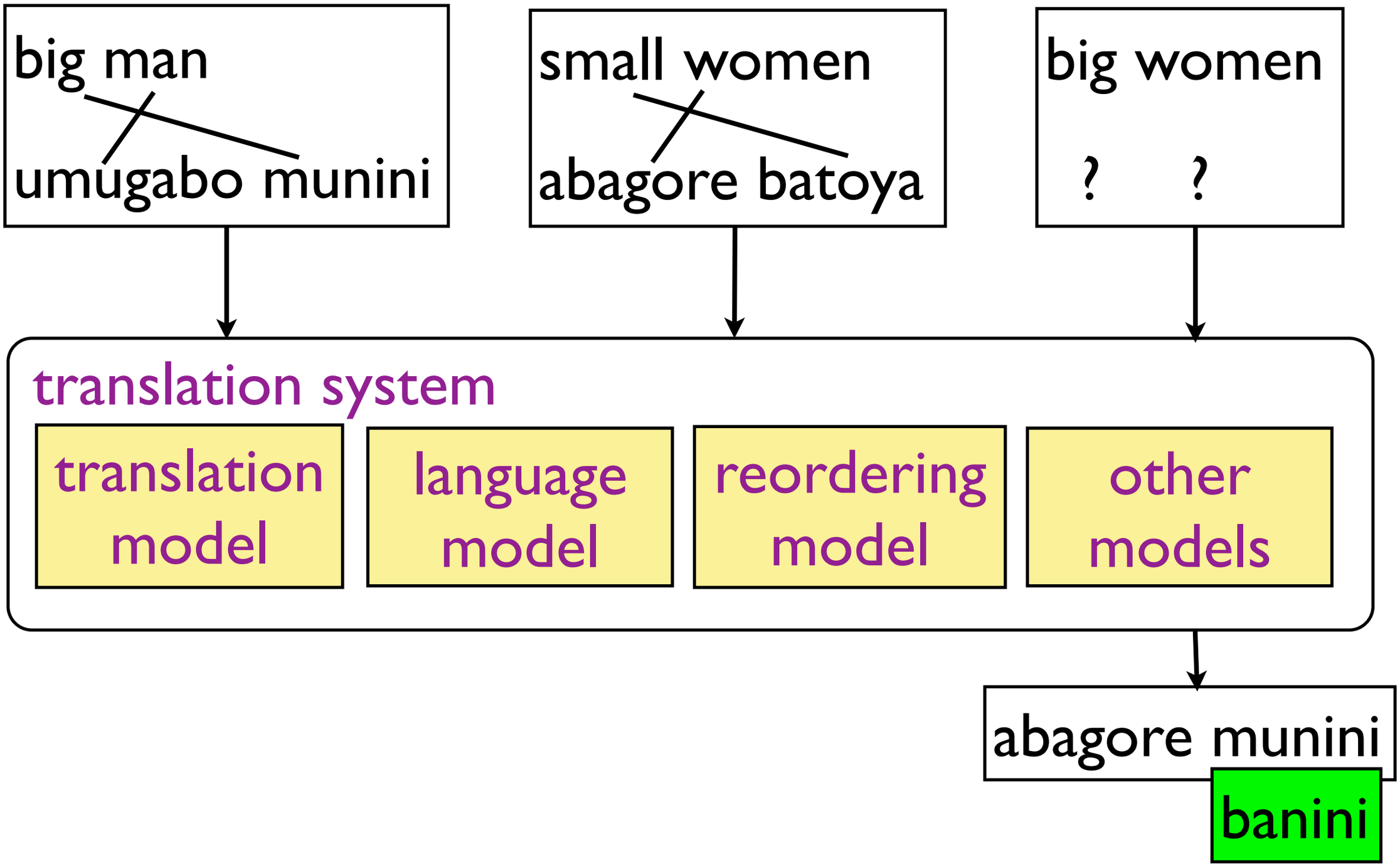
use morphological features + approximate inference

Problem #2: inability to generate unseen inflections.

Problem #3: n-gram LM fails to predict rare inflections.

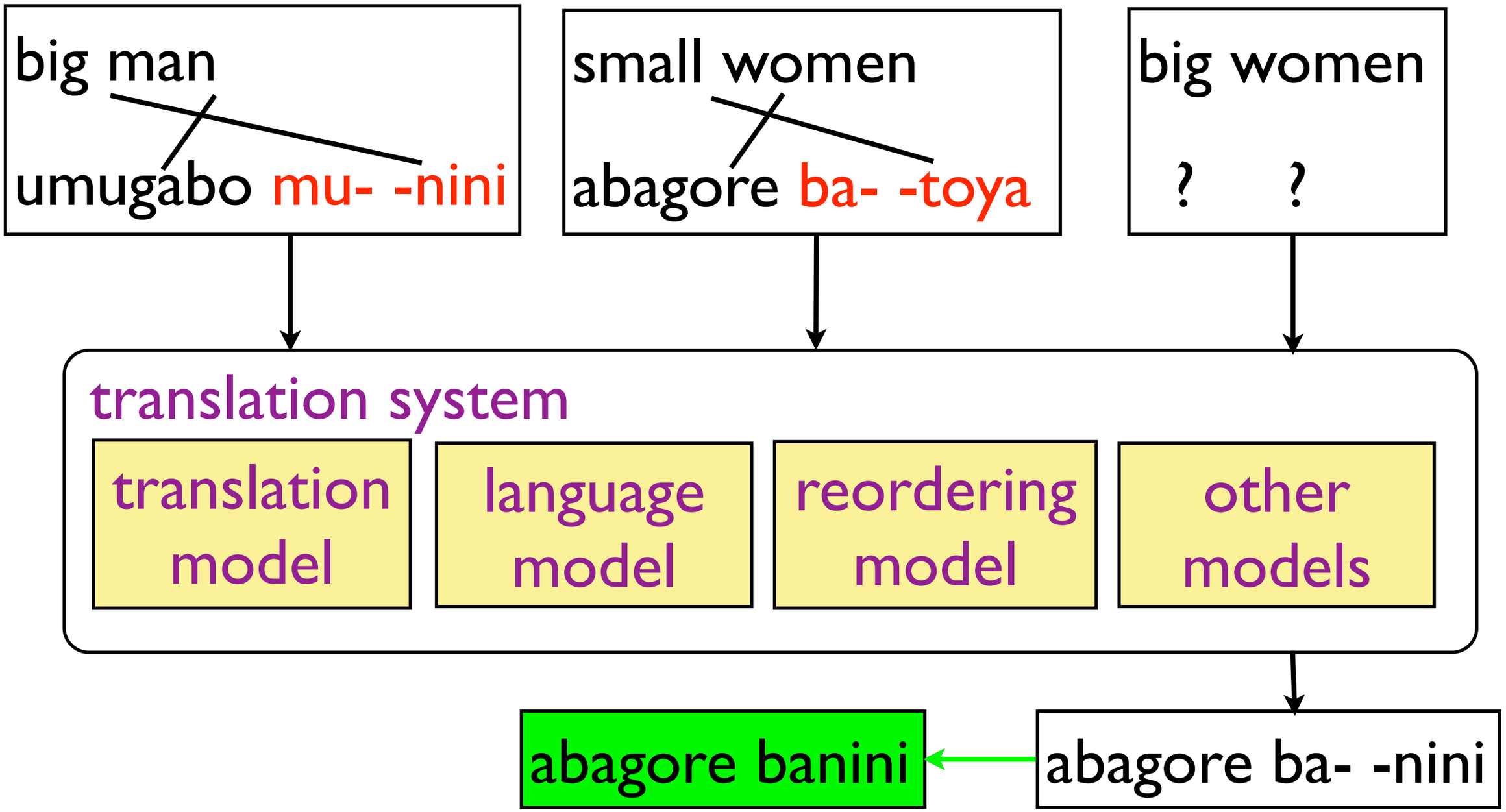
Problem #2: inability to generate unseen inflections.

default lexical representation



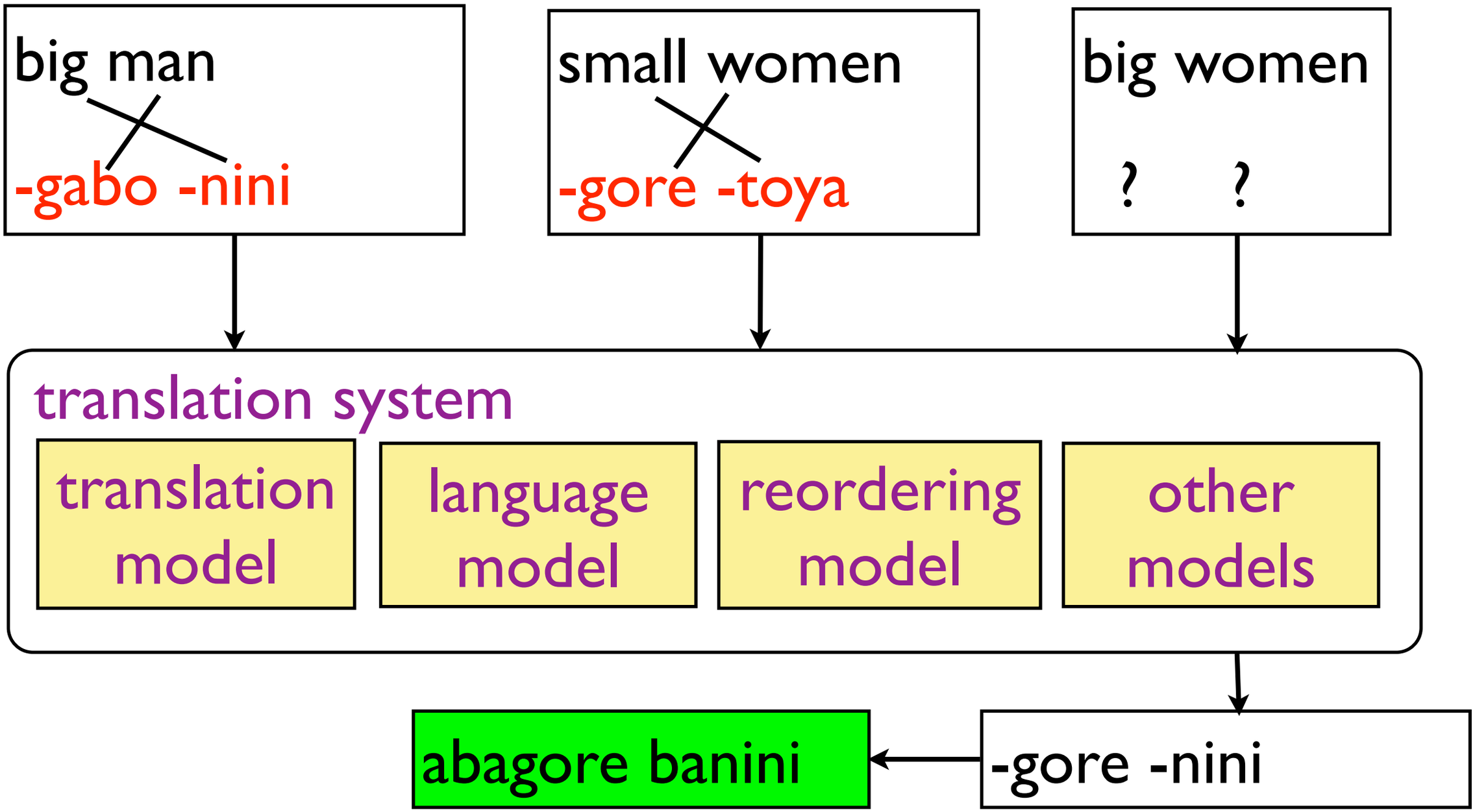
Problem #2: inability to generate unseen inflections.

alternative lexical representation (a)



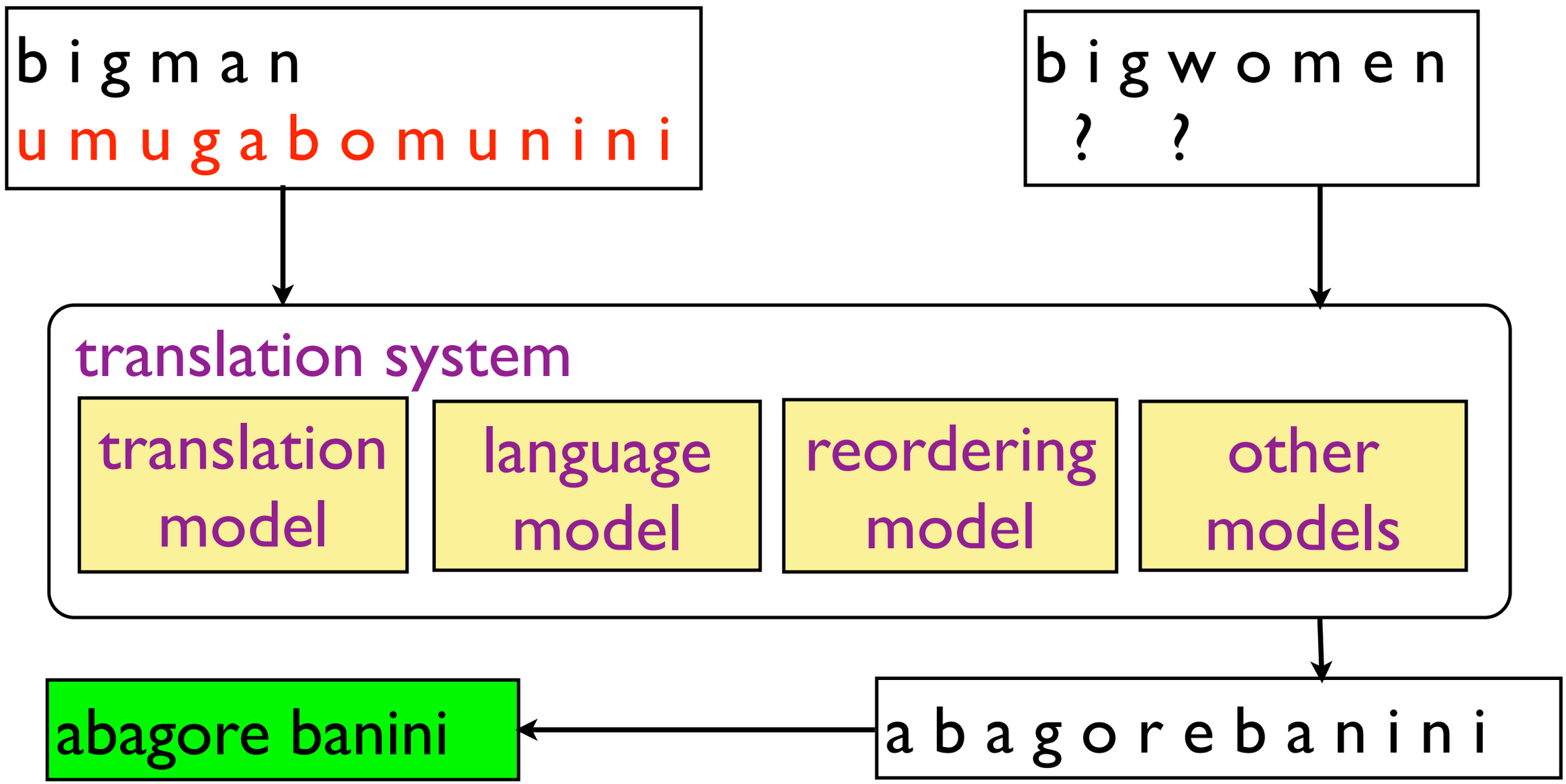
Problem #2: inability to generate unseen inflections.

alternative lexical representation (b)



Problem #2: inability to generate unseen inflections.

alternative lexical representation (c)



Problem #2: inability to generate unseen inflections.

alternative lexical representations

- some proposed representations:
  - affix transformations Habash and Sadat 2006, Al-Haj and Lavie 2010
  - artificially inflected English words Oflazer and El-Kahlout 2007
  - each character is a separate token Xu et al. 2004
- problems:
  - strong assumptions
  - subset of useful representations

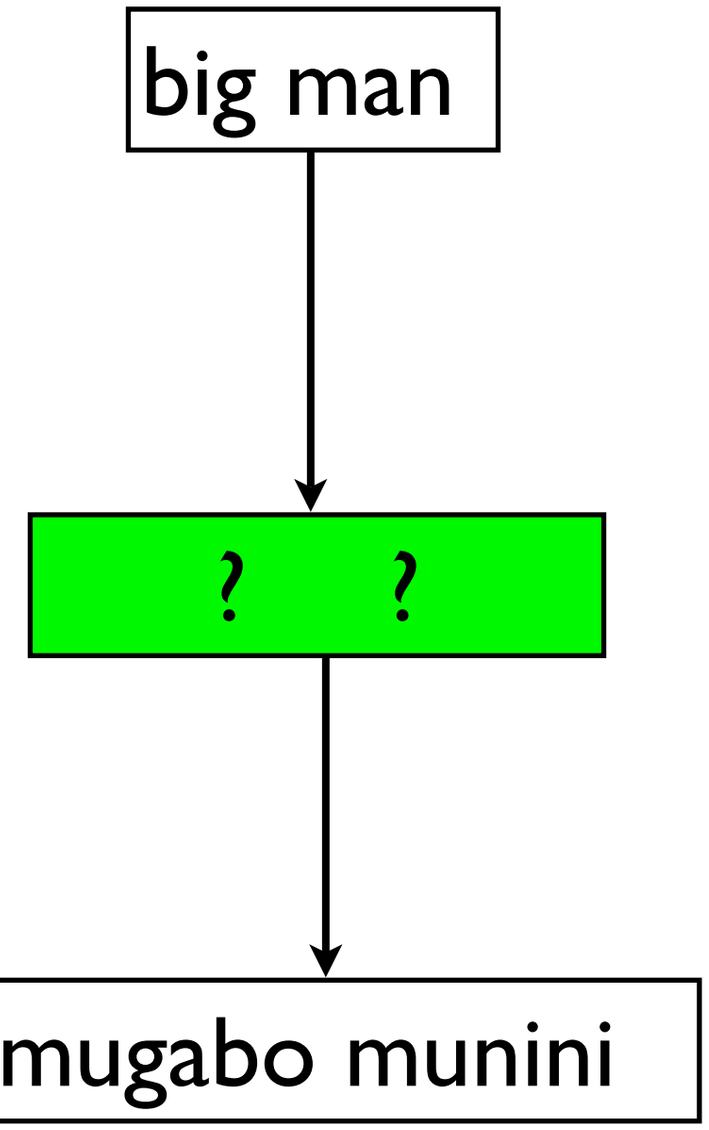
Problem #2: inability to generate unseen inflections.

representation variables

English

translation-friendly lexical representation

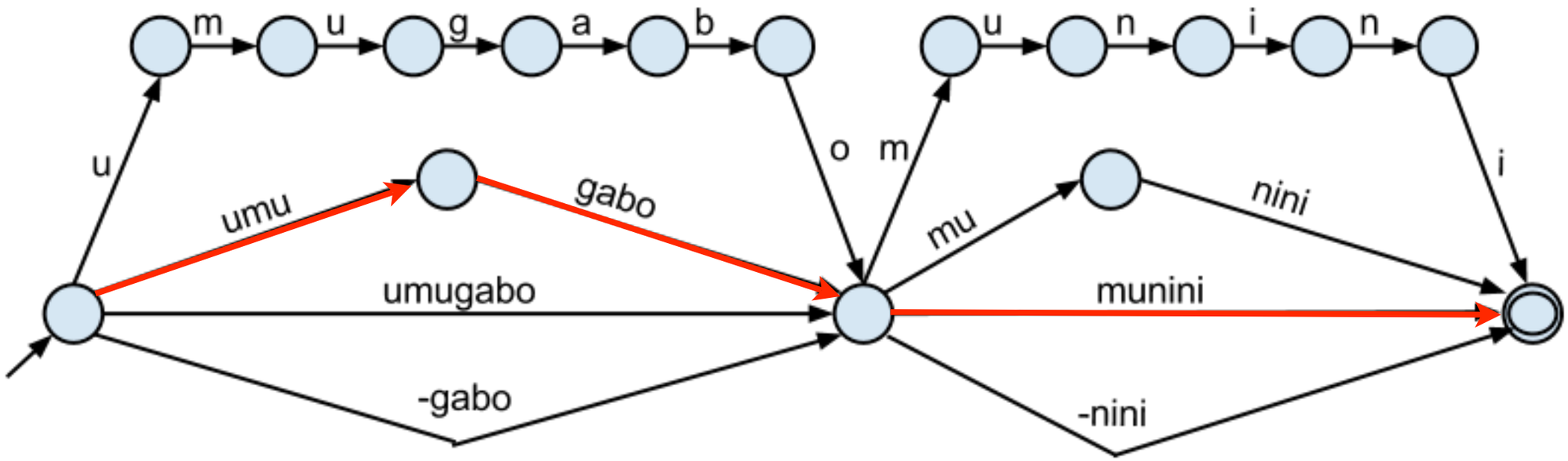
Kinyarwanda



possible values
umugabo mu nini
umu gabo munini
umu gabo mu nini
u m u g a b o m u n i n i
umugabo munini
gabo nini

Problem #2: inability to generate unseen inflections.

representation variables



- p(a,t|s)
- p(a,**r**,t | s) **r**epresentation
- approximate inference

road map

Problem #1: word alignment errors for rare inflections.

→ use morphological features + approximate inference

Problem #2: inability to generate unseen inflections.

→ learn translation-friendly lexical representations

Problem #3: n-gram LM fails to predict rare inflections.

Problem #3: n-gram LM fails to predict correct inflections.

n-gram LM favors popular sequences

big man
umu- -gabo mu- -nini

small women
aba- -gore ba- -toya

big women
? ?

translation system

translation model

language model

reordering model

other models

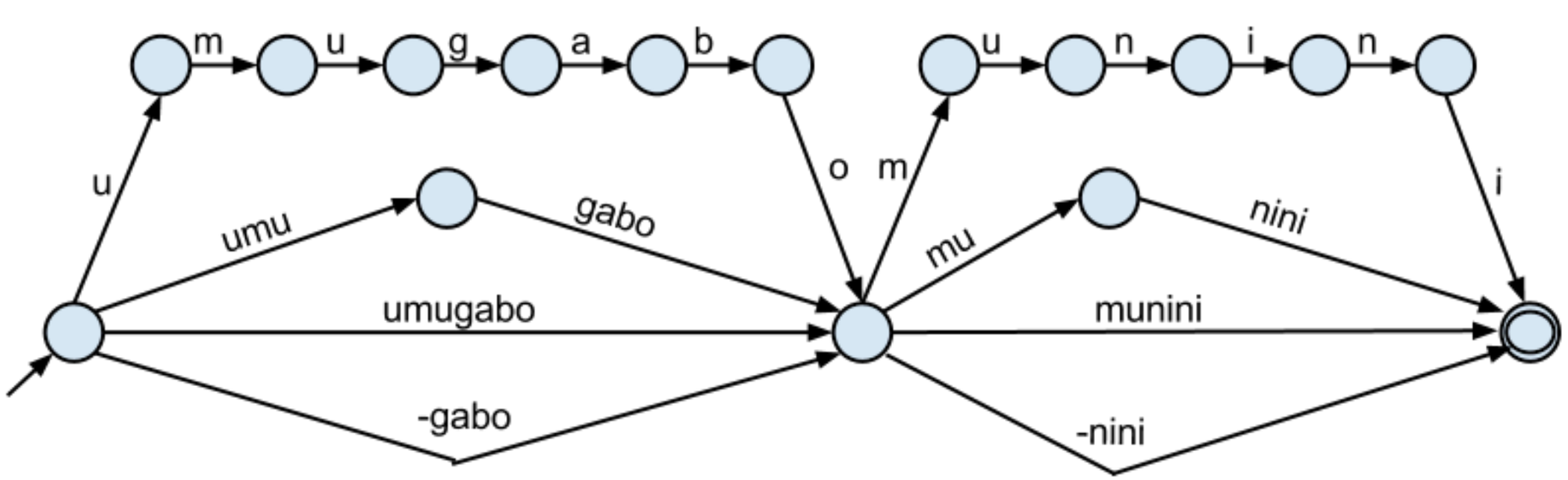
abagore munini

aba- -gore mu- -nini

banini

Problem #3: n-gram LM fails to predict correct inflections.

morphologically aware LM



road map

Problem #1: word alignment errors for rare inflections.

→ use morphological features + approximate inference

Problem #2: inability to generate unseen inflections.

→ learn translation-friendly lexical representations

Problem #3: n-gram LM fails to predict rare inflections.

→ morphologically-aware language model

models that use linguistic knowledge

inside translation model

questions?