Statistical Machine Translation without Parallel Data

Maryam Siahbani

Supervisor: Dr. Anoop Sarkar

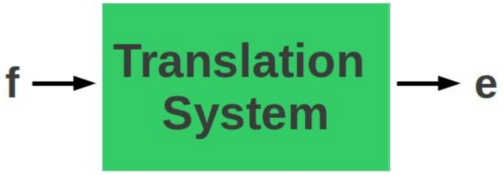
Natural Language Laboratory
Simon Fraser University

Statistical Machine Translation (SMT)

$$e = \arg\max_{e} \{ \Pr(e \mid f) \}$$

e: target sentence

f: source sentence

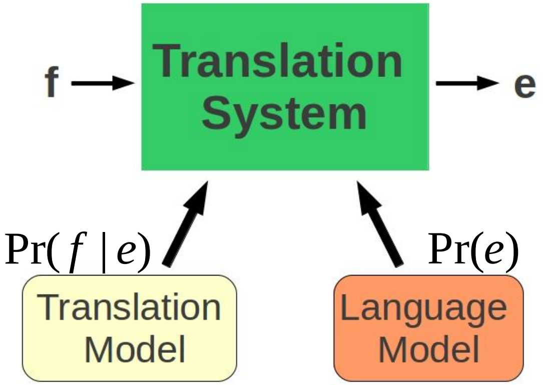


Statistical Machine Translation (SMT)

$$e = \underset{e}{\arg\max}\{\Pr(e|f)\}$$

e: target sentence

f: source sentence



Noisy channel

$$\Pr(e|f) \propto \Pr(e).\Pr(f|e)$$

Log-linear model

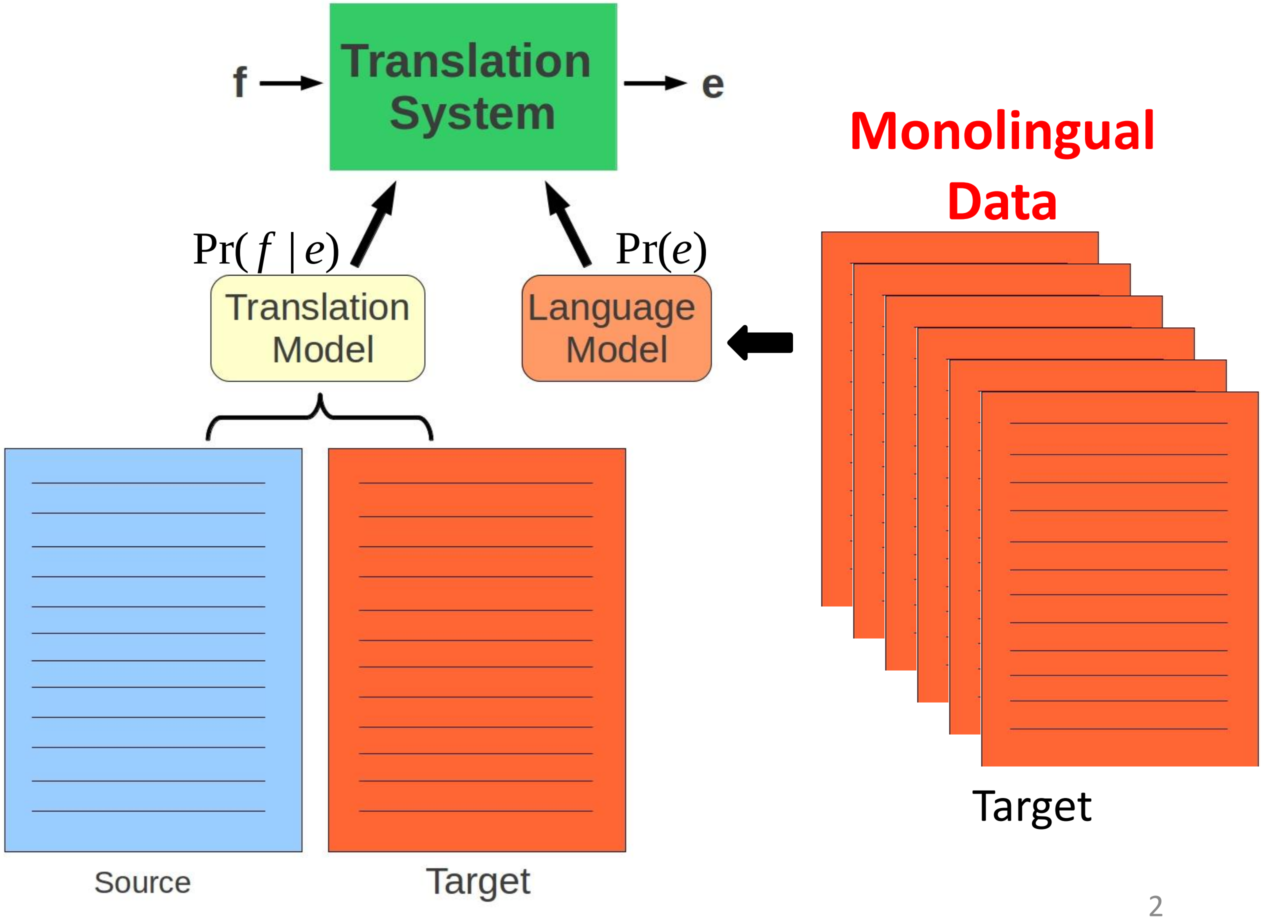
$$\Pr(e|f) \propto \exp \sum_i \lambda_i h_i(e,f)$$

Statistical Machine Translation (SMT)

$$e = \arg\max_{e} \{\Pr(e \mid f)\}$$

e: target sentence

f: source sentence



f → Translation System → e

$\Pr(f \mid e)$

Translation Model

$\Pr(e)$

Language Model

Monolingual Data

Target

Source

Target

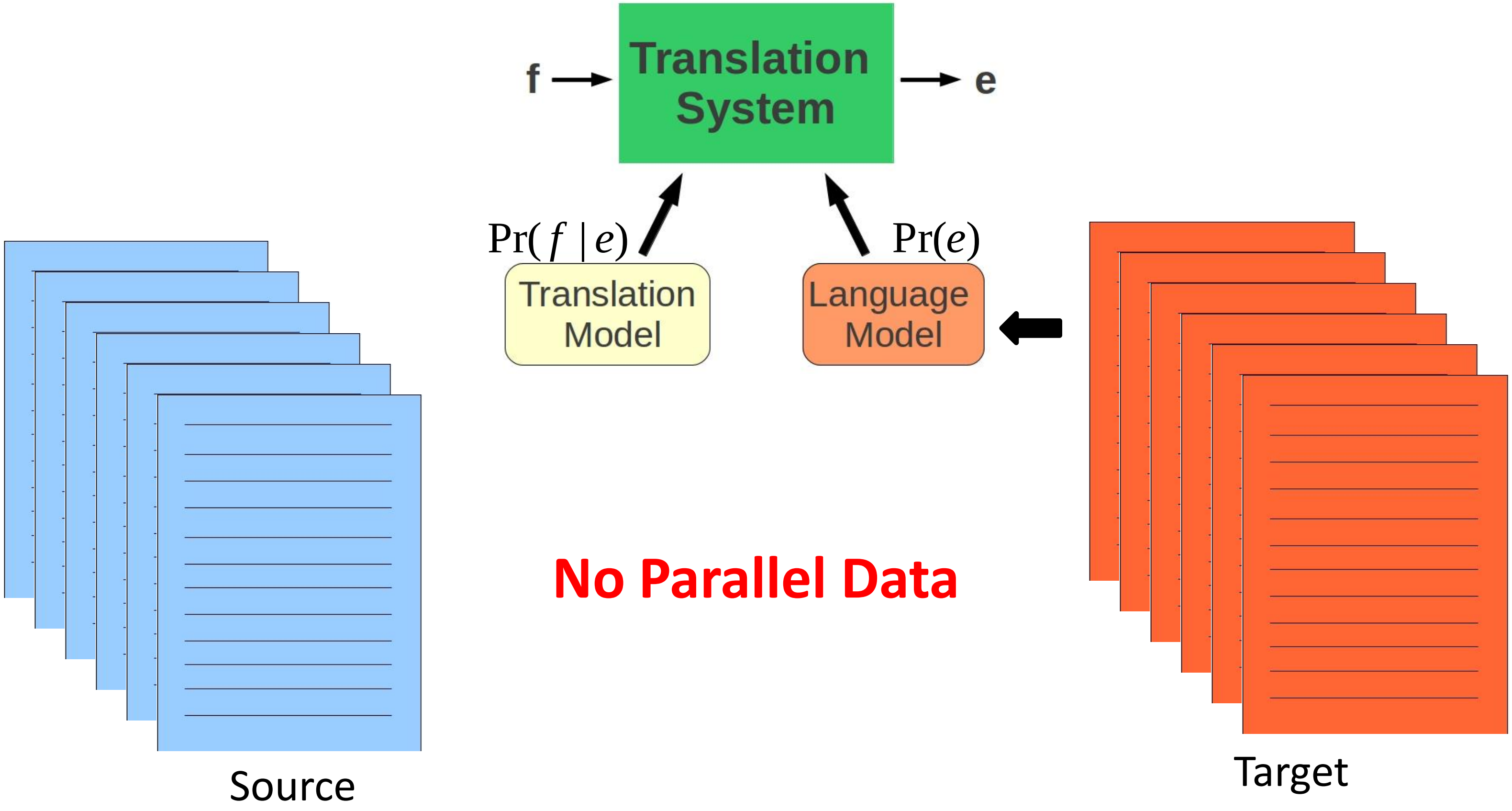
Parallel Data (or bilingual data)

Statistical Machine Translation (SMT)

$$e = \arg\max_{e}\{\Pr(e \mid f)\}$$

e: target sentence

f: source sentence



No Parallel Data

Methods

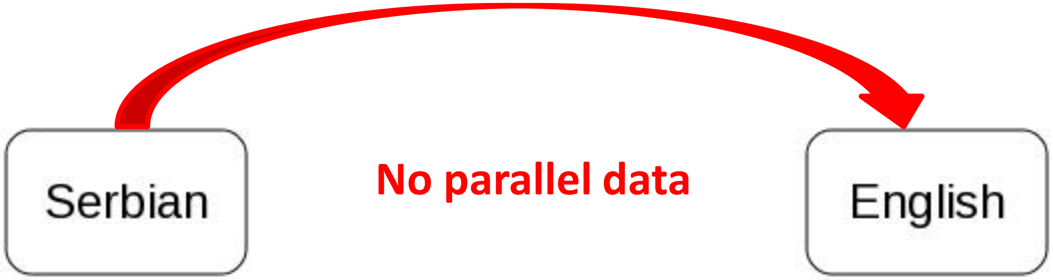
- Translation lexicon
  - ➢ Limited parallel resource
  - ➢ Monolingual corpora

- Translation system

Translation Lexicon Induction

Limited Parallel Data

Inducing Translation Lexicon: Bridge language



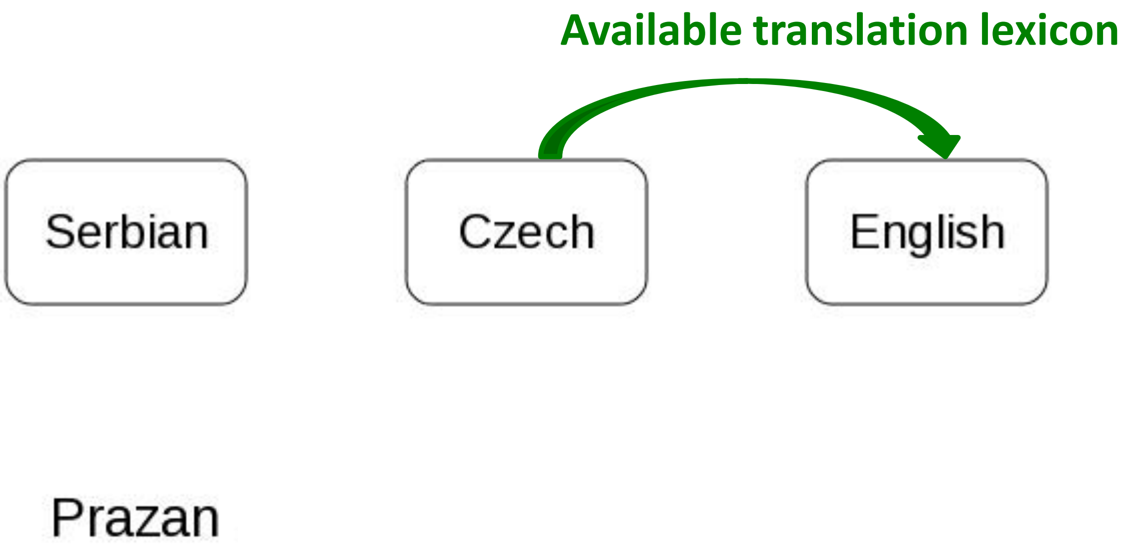
Serbian

No parallel data

English

Prazan

Inducing Translation Lexicon: Bridge language



Available translation lexicon

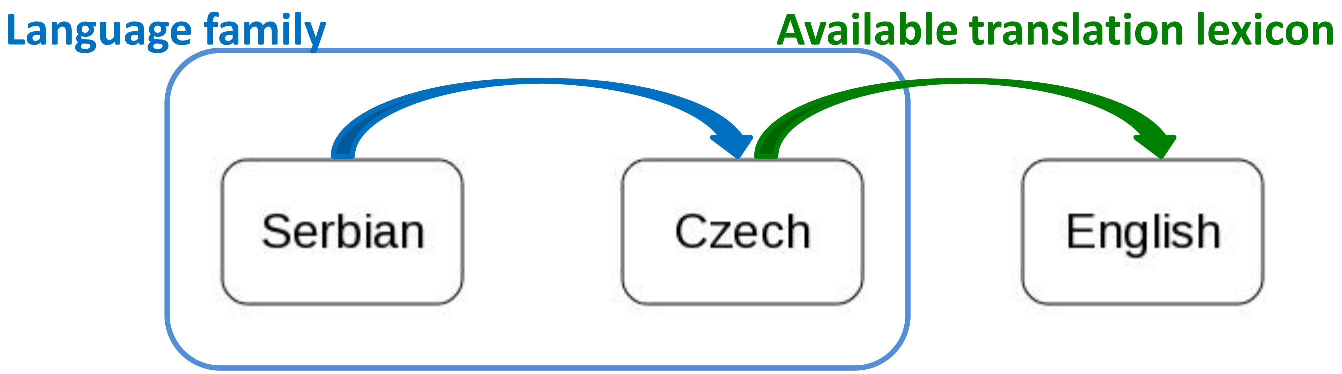
Serbian

Czech

English

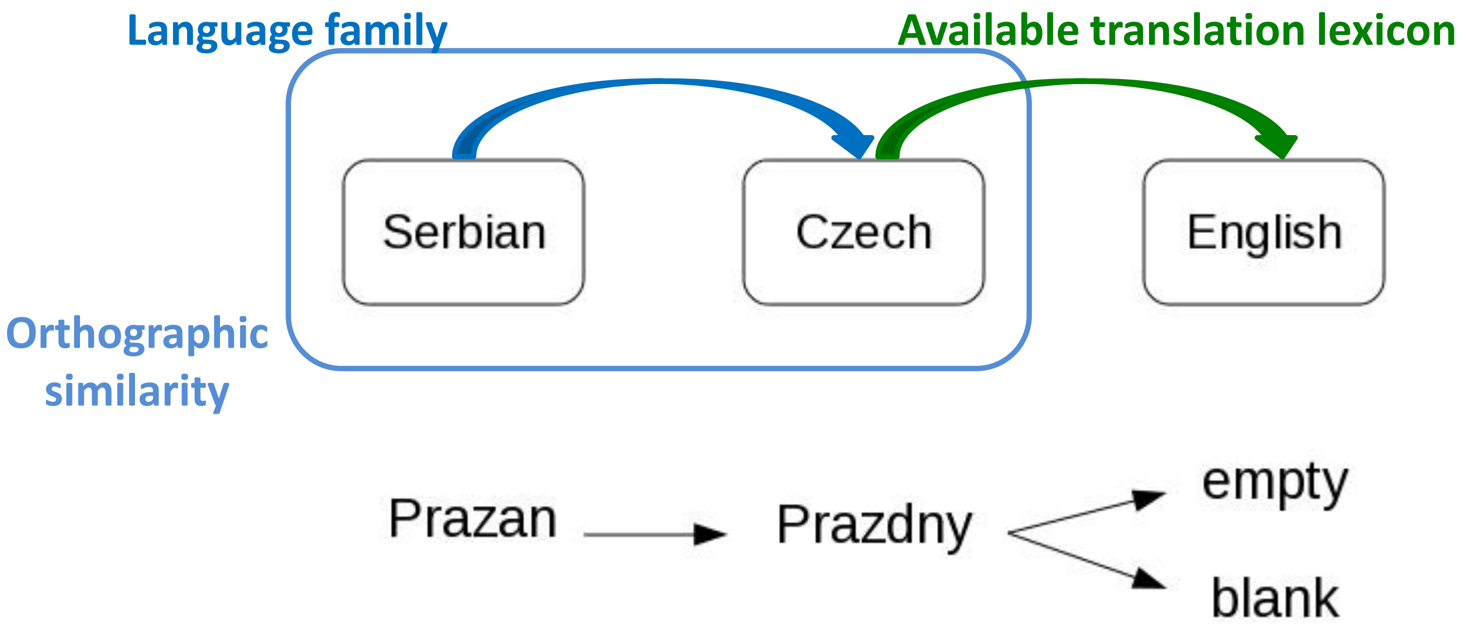
Prazan

Inducing Translation Lexicon: Bridge language



Prazan → Prazdny

Inducing Translation Lexicon: Bridge language



Language family

Available translation lexicon

Orthographic similarity

Serbian

Czech

English

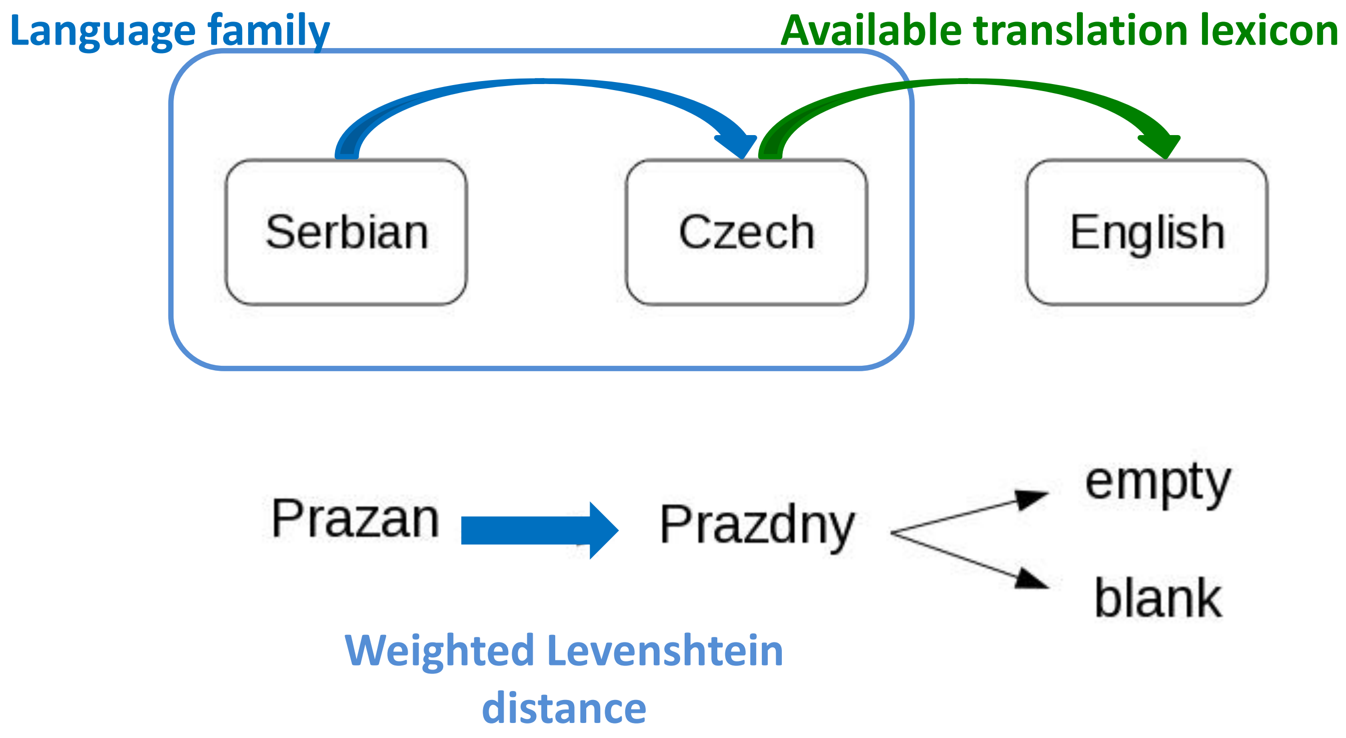
Prazan

Prazdny

empty

blank

Inducing Translation Lexicon: Bridge language (Mann and Yarowsky, 2001)



Operations in Levenshtein:

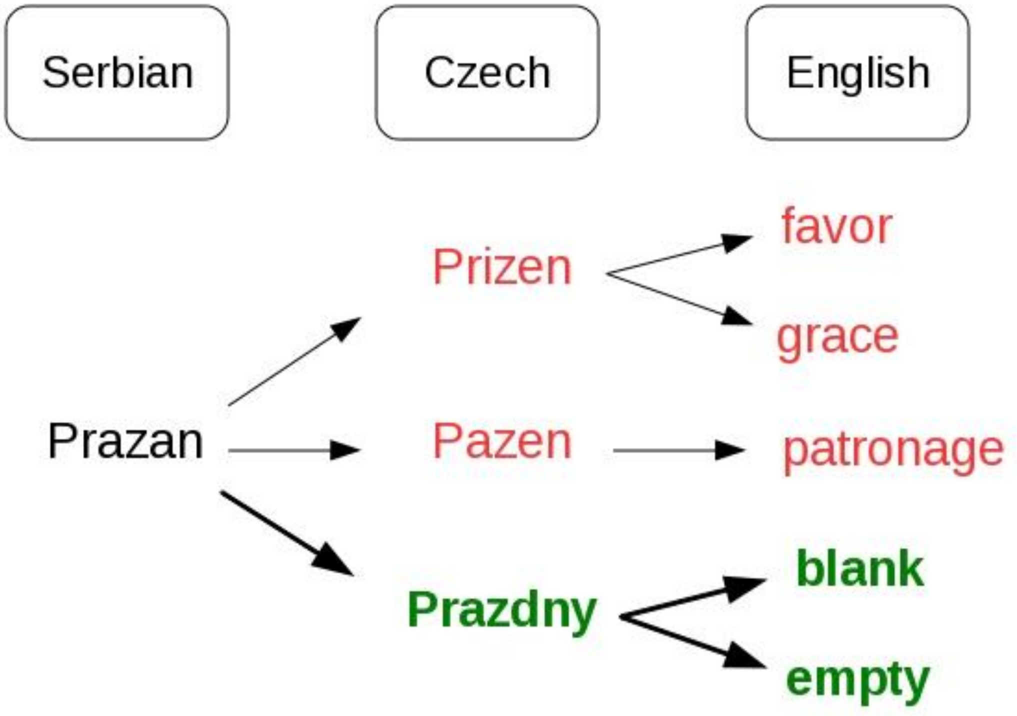
Substitution: Otpor -> Odpor, cost (t,d)

Insertion: Pravo -> Vprovo , cost(ϵ, v)

Deletion: Chvala -> Hvaliti , cost (c, ϵ)

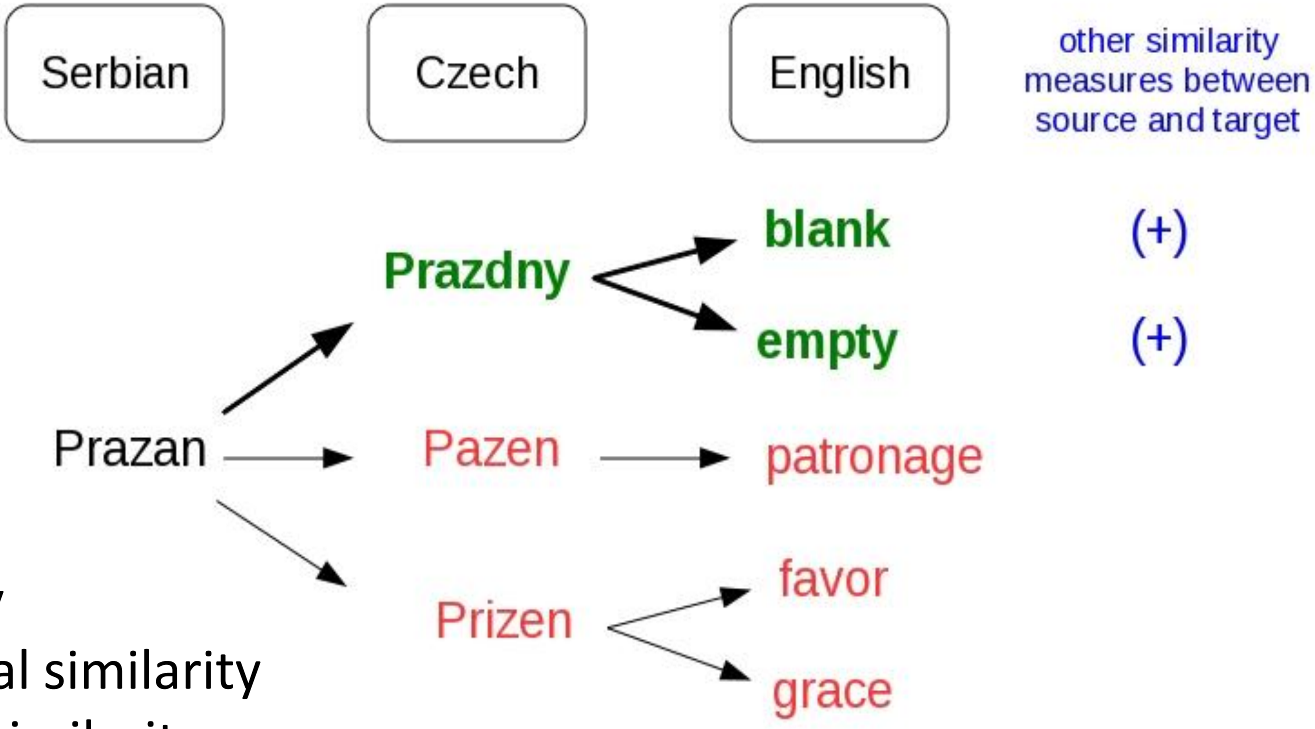
Inducing Translation Lexicon: Bridge language

- False cognates



Inducing Translation Lexicon: Bridge language (Schafer and Yarowsky 2002)

- Additional similarity measures



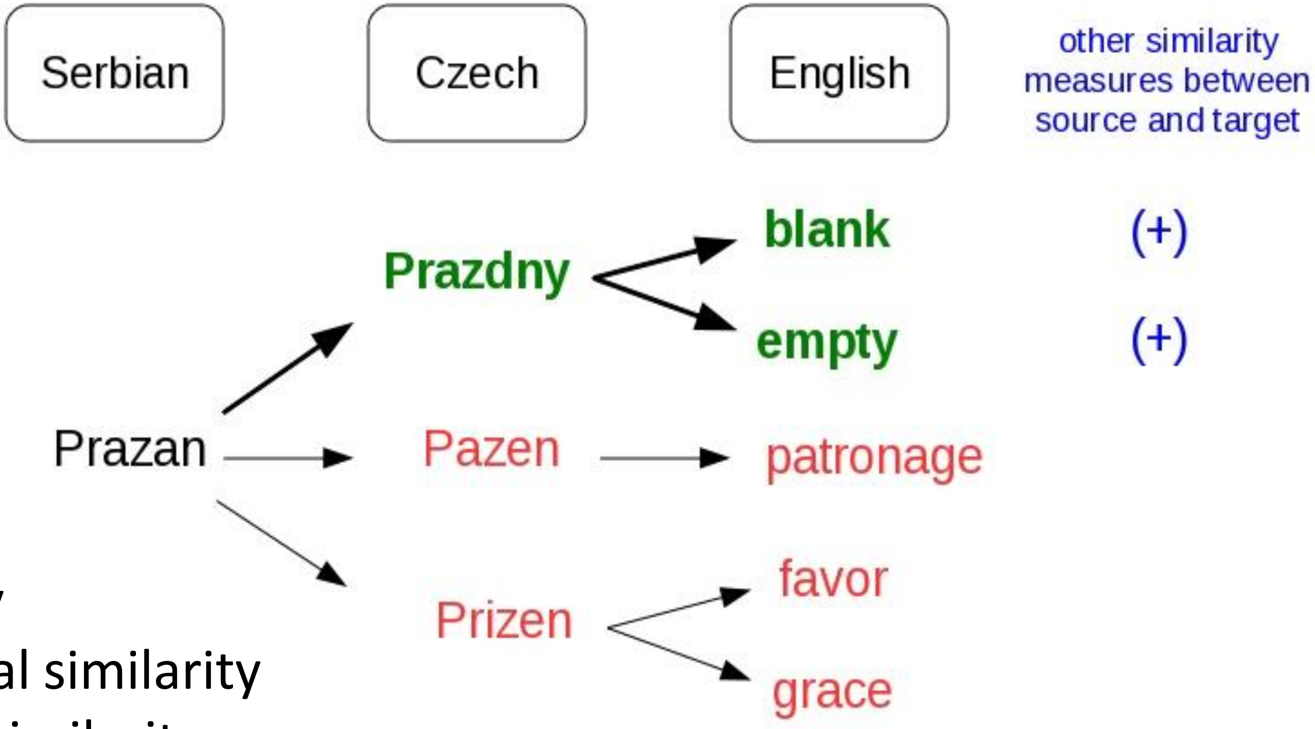
- Context similarity
- Date distributional similarity
- Word frequency similarity

Inducing Translation Lexicon: Bridge language (Schafer and Yarowsky 2002)

- Additional similarity measures

9% Improvement

Accuracy measured against a gold translation lexicon



- Context similarity
- Date distributional similarity
- Word frequency similarity

Inducing Translation Lexicon: Using Seed Lexicon

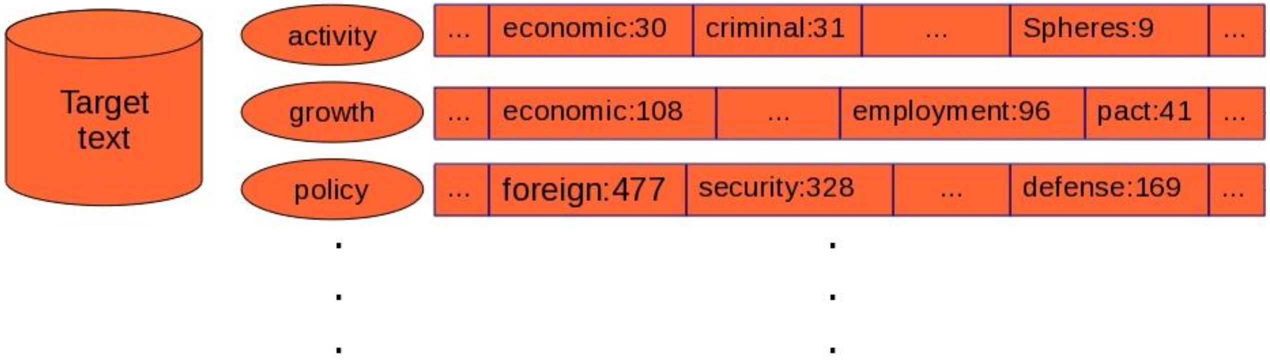
- Given: an initial small lexicon between source and target

- Goal: a translation lexicon between S and T

- Methods: extending the seed lexicon
  - Context similarity

Inducing Translation Lexicon: Using Seed Lexicon (Context similarity)

- (Rapp 95),(Rapp 99),(Fung & Yee 98)



Target text

activity ... economic:30 criminal:31 ... Spheres:9 ...

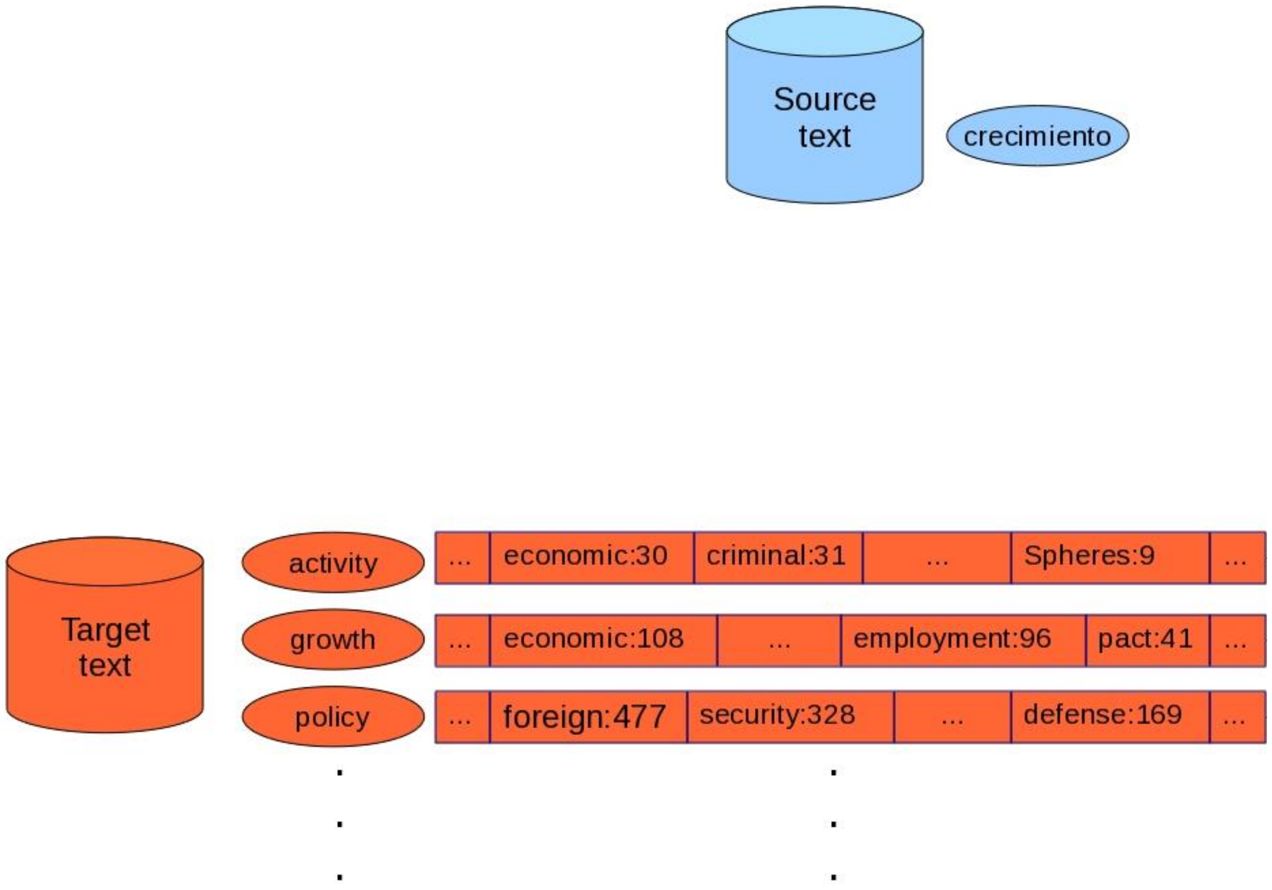
growth ... economic:108 ... employment:96 pact:41 ...

policy ... foreign:477 security:328 ... defense:169 ...

.
.
.

Inducing Translation Lexicon: Using Seed Lexicon (Context similarity)

- (Rapp 95),(Rapp 99),(Fung & Yee 98)



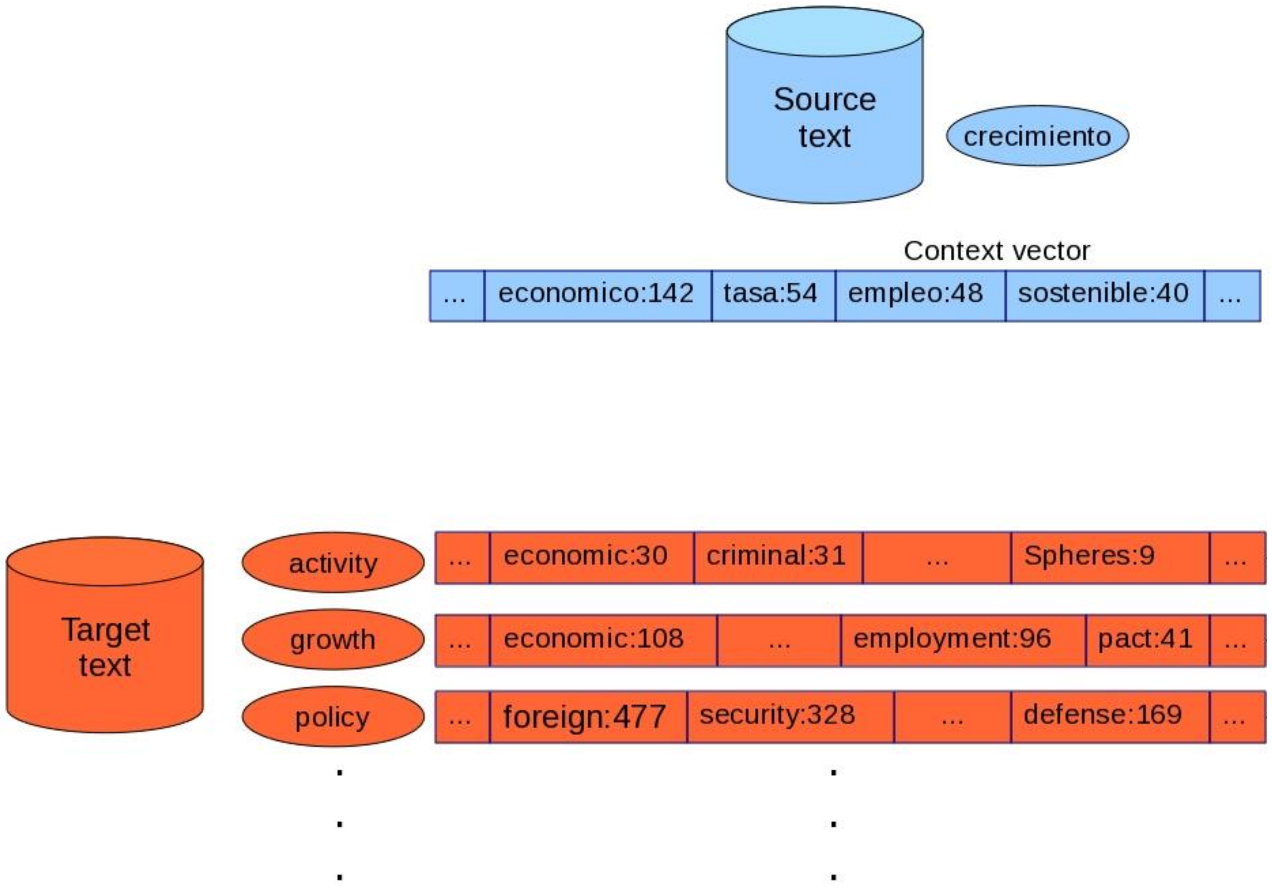
activity: … | economic:30 | criminal:31 | … | Spheres:9 | …

growth: … | economic:108 | … | employment:96 | pact:41 | …

policy: … | foreign:477 | security:328 | … | defense:169 | …

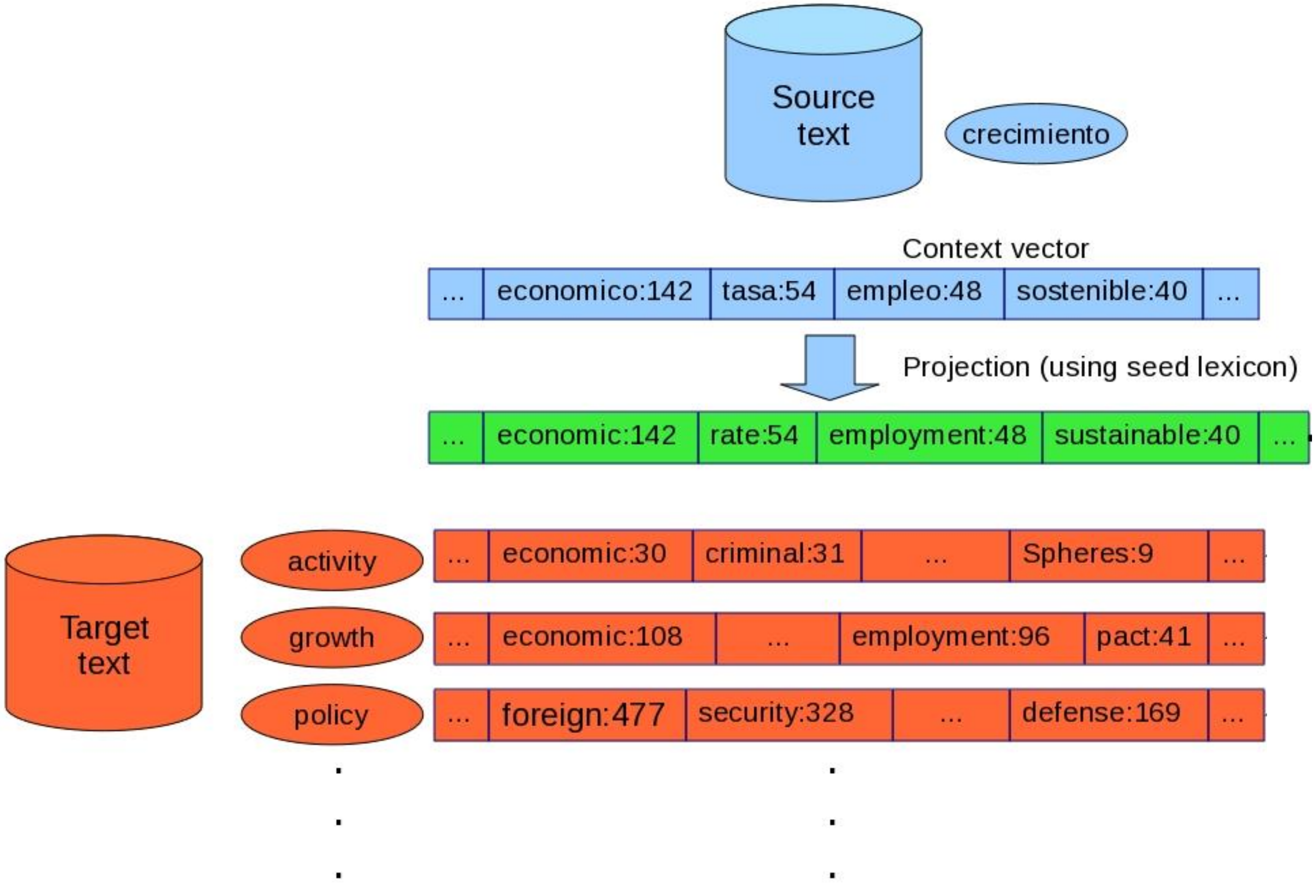
Inducing Translation Lexicon: Using Seed Lexicon (Context similarity)

- (Rapp 95),(Rapp 99),(Fung & Yee 98)


Source text
crecimiento
Context vector
...
economico:142
tasa:54
empleo:48
sostenible:40
...
Target text
activity
... economic:30 criminal:31 ... Spheres:9 ...
growth
... economic:108 ... employment:96 pact:41 ...
policy
... foreign:477 security:328 ... defense:169 ...
.
.
.

Inducing Translation Lexicon: Using Seed Lexicon (Context similarity)

- (Rapp 95),(Rapp 99),(Fung & Yee 98)



Source text

crecimiento

Context vector

… | economico:142 | tasa:54 | empleo:48 | sostenible:40 | …

Projection (using seed lexicon)

… | economic:142 | rate:54 | employment:48 | sustainable:40 | …

Target text

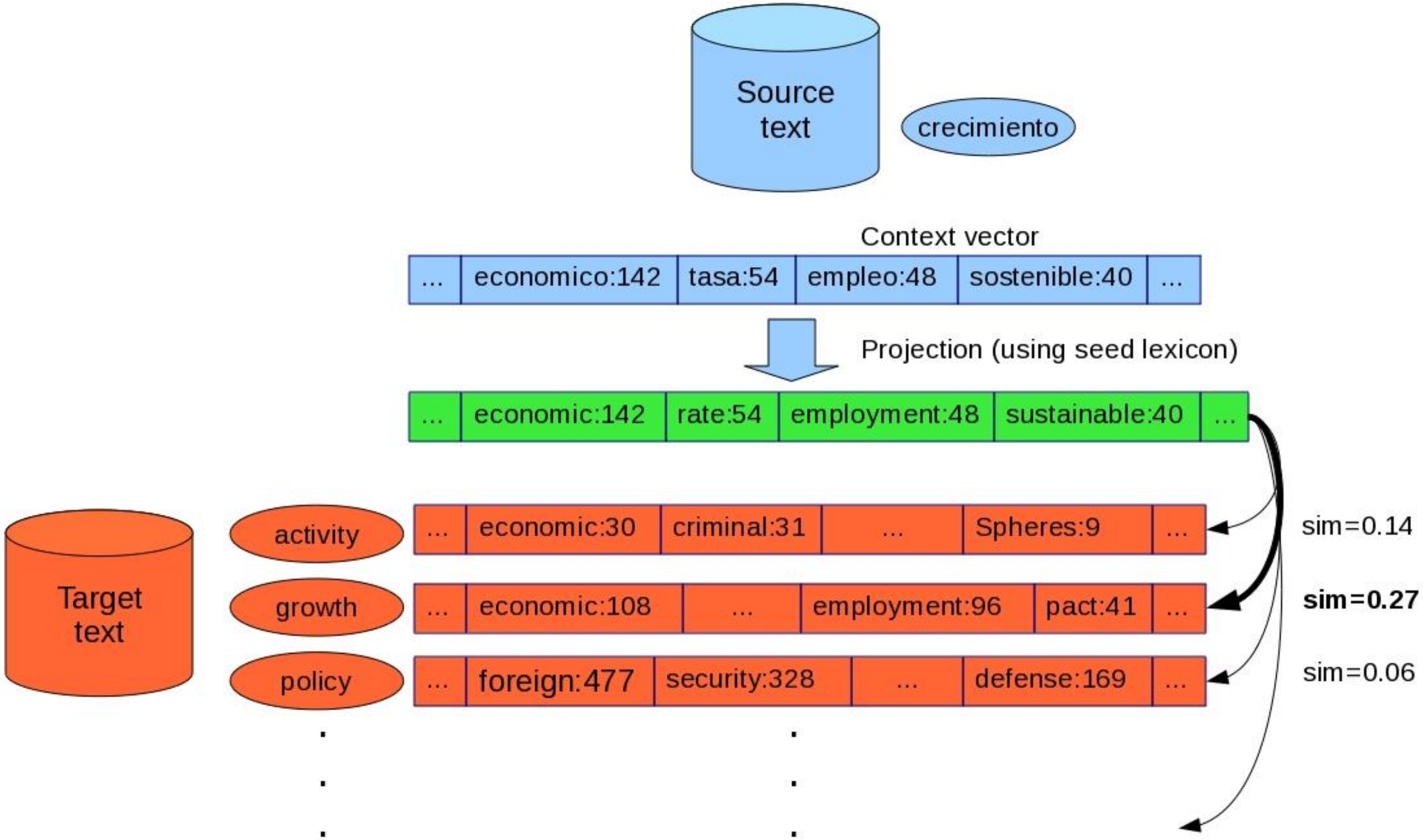
activity: … | economic:30 | criminal:31 | … | Spheres:9 | …

growth: … | economic:108 | … | employment:96 | pact:41 | …

policy: … | foreign:477 | security:328 | … | defense:169 | …

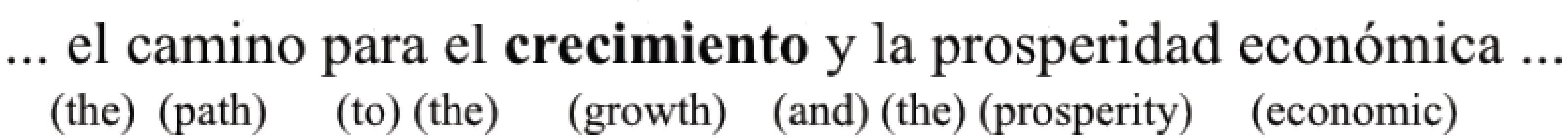
Inducing Translation Lexicon: Using Seed Lexicon (Context similarity)

- (Rapp 95),(Rapp 99),(Fung & Yee 98)


Source text
crecimiento
Context vector
... | economico:142 | tasa:54 | empleo:48 | sostenible:40 | ...
Projection (using seed lexicon)
... | economic:142 | rate:54 | employment:48 | sustainable:40 | ...
Target text
activity
... | economic:30 | criminal:31 | ... | Spheres:9 | ...
sim=0.14
growth
... | economic:108 | ... | employment:96 | pact:41 | ...
sim=0.27
policy
... | foreign:477 | security:328 | ... | defense:169 | ...
sim=0.06

Inducing Translation Lexicon: Using Seed Lexicon (Context similarity)

- Modeling the context
  - Window of fixed size (Rapp 95),(Rapp 99),(Fung & Yee 98)

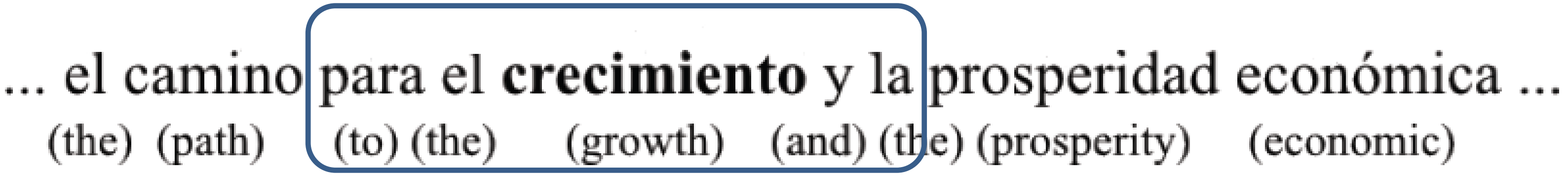


... el camino para el **crecimiento** y la prosperidad económica ...

(the) (path) (to) (the) (growth) (and) (the) (prosperity) (economic)

Inducing Translation Lexicon: Using Seed Lexicon (Context similarity)

- Modeling the context
  - Window of fixed size (Rapp 95),(Rapp 99),(Fung & Yee 98)



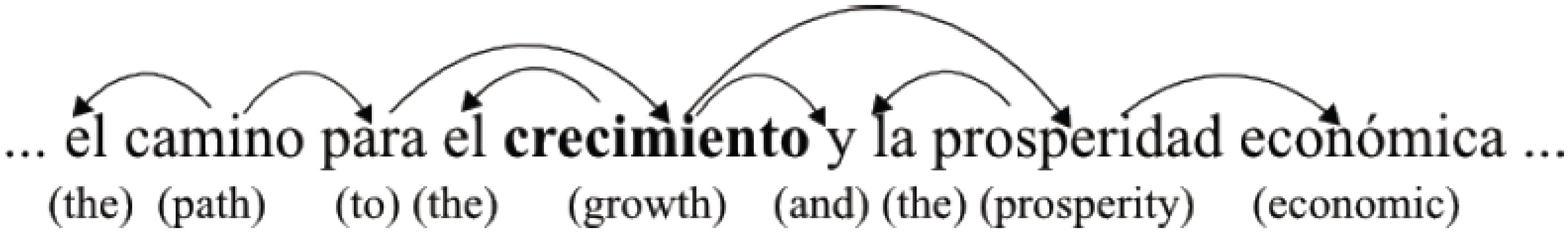
... el camino para el **crecimiento** y la prosperidad económica ...

(the) (path) (to) (the) (growth) (and) (the) (prosperity) (economic)

Position	Adjacent context
-2	para
-1	el
+1	y
+2	la

Inducing Translation Lexicon: Using Seed Lexicon (Context similarity)

- Modeling the context
  - Window of fixed size (Rapp 95),(Rapp 99),(Fung & Yee 98)
  - Using dependency trees (Garera et al., 2009)



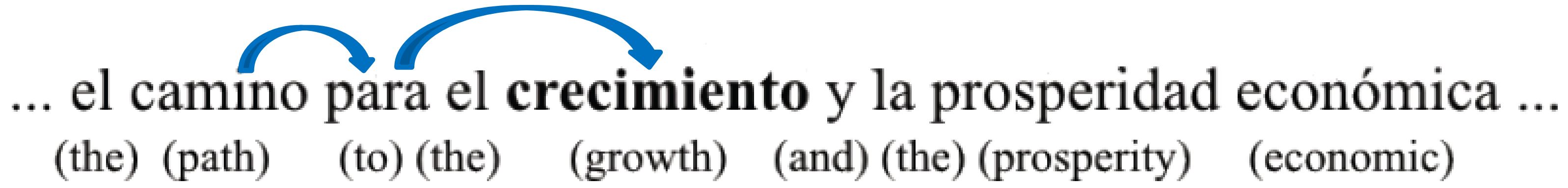
... el camino para el **crecimiento** y la prosperidad económica ...

(the) (path) (to) (the) (growth) (and) (the) (prosperity) (economic)

Position	Adjacent context	Dependency context
-2	para	
-1	el	
+1	y	
+2	la	

Inducing Translation Lexicon: Using Seed Lexicon (Context similarity)

- Modeling the context
  - Window of fixed size (Rapp 95),(Rapp 99),(Fung & Yee 98)
  - Using dependency trees (Garera et al., 2009)



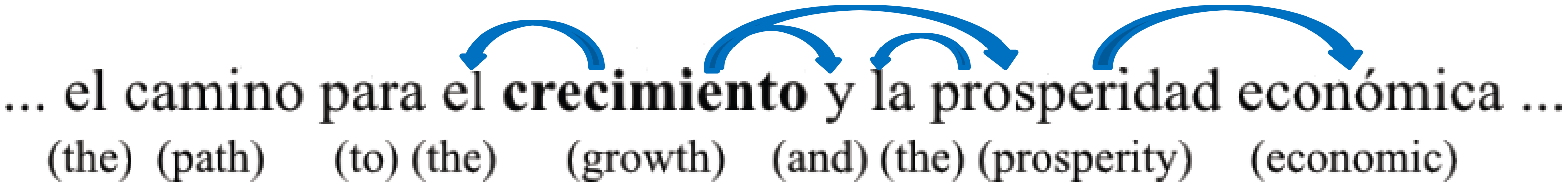
... el camino para el **crecimiento** y la prosperidad económica ...

(the) (path) (to) (the) (growth) (and) (the) (prosperity) (economic)

Position	Adjacent context	Dependency context
-2	para	camino
-1	el	para
+1	y	
+2	la	

Inducing Translation Lexicon: Using Seed Lexicon (Context similarity)

- Modeling the context
  - Window of fixed size (Rapp 95),(Rapp 99),(Fung & Yee 98)
  - Using dependency trees (Garera et al., 2009)



... el camino para el **crecimiento** y la prosperidad económica ...

(the) (path) (to) (the) (growth) (and) (the) (prosperity) (economic)

Position	Adjacent context	Dependency context
-2	para	camino
-1	el	para
+1	y	prosperidad, y, el
+2	la	la, economica

← **Dynamic context size**

Inducing Translation Lexicon: Using Seed Lexicon (Context similarity)

- Modeling the context
  - Window of fixed size (Rapp 95),(Rapp 99),(Fung & Yee 98)
  - Using dependency trees (Garera et al., 2009)

... el camino para
(the) (path) (to) (
dad económica ...
ty) (economic)

8% Improvement over adjacent context

Position	Adjacent context	
-2	para	
-1	el	
+1	y	prosperidad, y, el
+2	la	la, economica

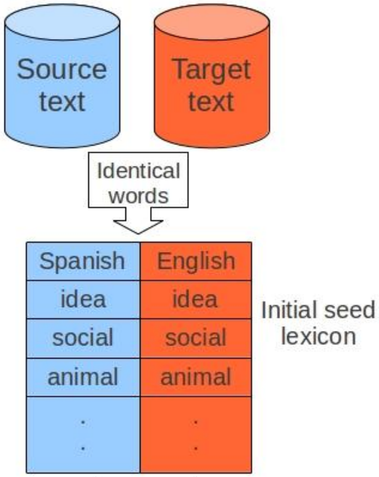
Spanish-English (1000 most frequent words)

context size

No Parallel Data

Inducing Translation Lexicon: Monolingual Data (Similarity measures)

- Similarities in monolingual corpora (Koehn & Knight 2002)
    - Identical words



Source text

Target text

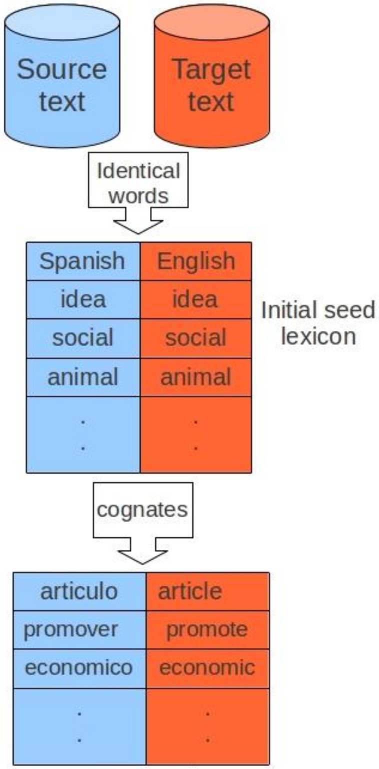
Identical words

Spanish	English
idea	idea
social	social
animal	animal
. .	. .

Initial seed lexicon

Inducing Translation Lexicon: Monolingual Data (Similarity measures)

- Similarities in monolingual corpora (Koehn & Knight 2002)
  - Identical words
  - Orthographic similarity (cognates)



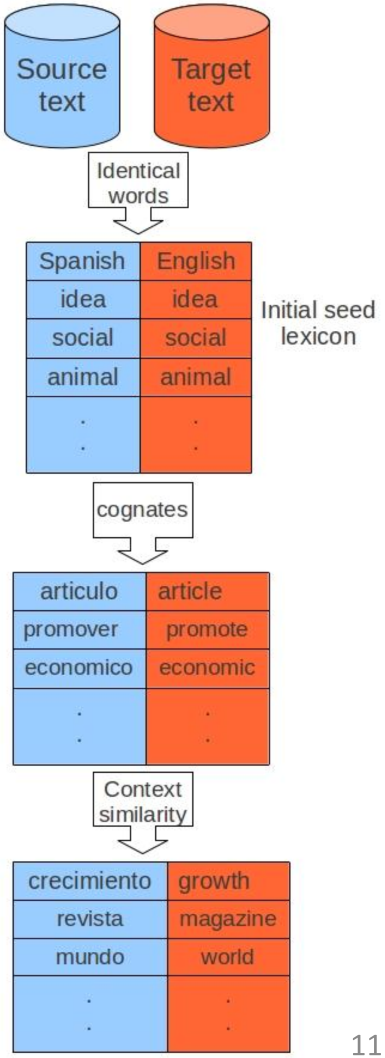
Spanish	English
idea	idea
social	social
animal	animal
. .	. .

Initial seed lexicon

articulo	article
promover	promote
economico	economic
. .	. .

Inducing Translation Lexicon: Monolingual Data (Similarity measures)

- Similarities in monolingual corpora (Koehn & Knight 2002)
  - Identical words
  - Orthographic similarity (cognates)
  - Context similarity



Source text

Target text

Identical words

Spanish	English
idea	idea
social	social
animal	animal
⋮	⋮

Initial seed lexicon

cognates

articulo	article
promover	promote
economico	economic
⋮	⋮

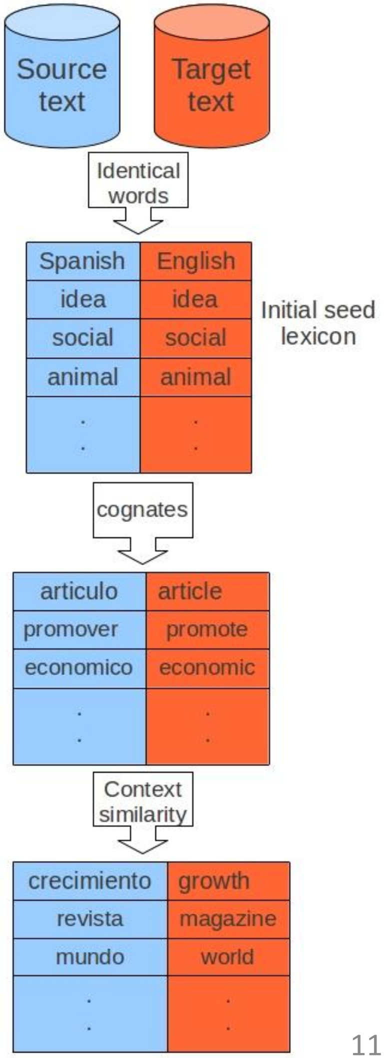
Context similarity

crecimiento	growth
revista	magazine
mundo	world
⋮	⋮

Inducing Translation Lexicon: Monolingual Data (Similarity measures)

- Similarities in monolingual corpora (Koehn & Knight 2002)
  - Identical words
  - Orthographic similarity (cognates)
  - Context similarity

German-English
1000 most frequent nouns



Spanish	English
idea	idea
social	social
animal	animal
. .	. .

articulo	article
promover	promote
economico	economic
. .	. .

crecimiento	growth
revista	magazine
mundo	world
. .	. .

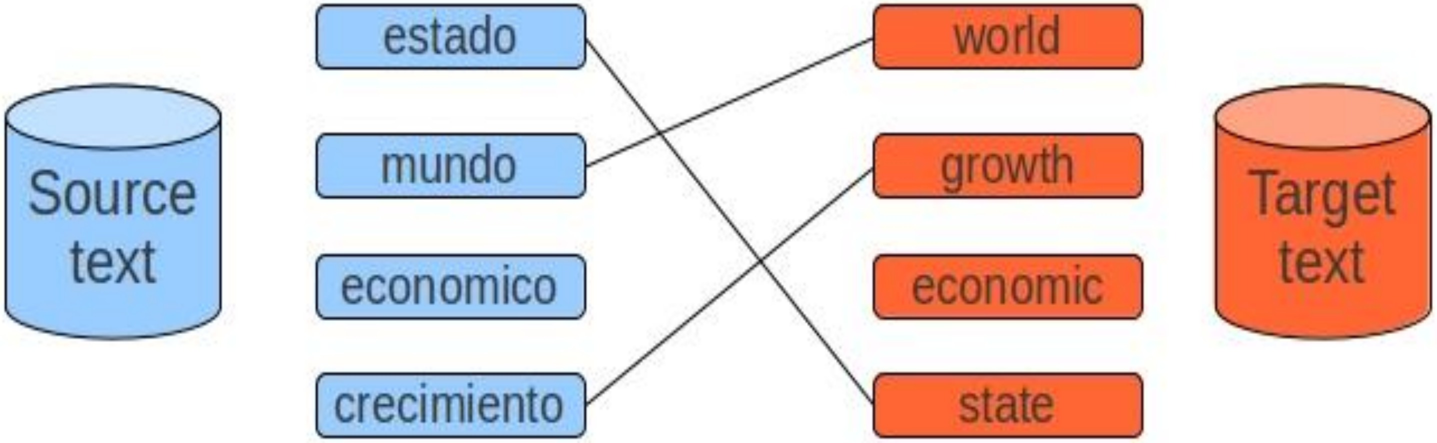
~ 16%

25.4%

39%

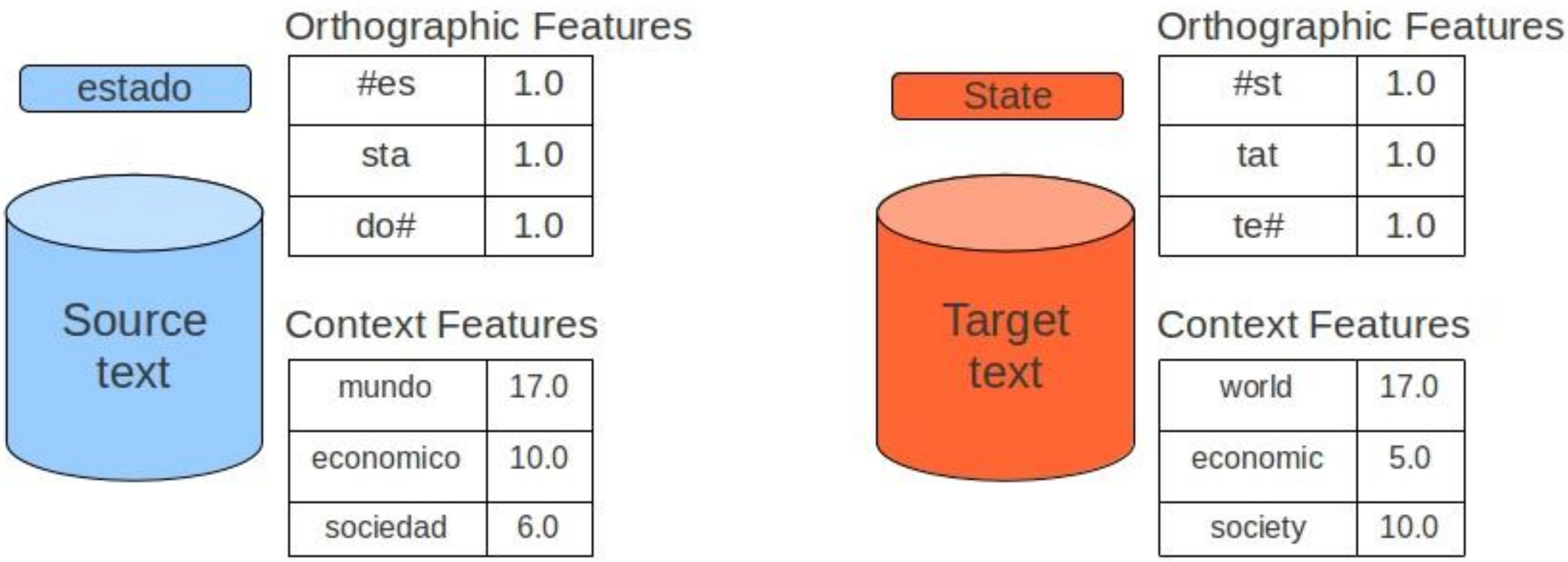
Inducing Translation Lexicon:
No Parallel Data (Model Translation Lexicon as a Mapping)

- A generative model for matching between S and T (Haghighi et al., 2008)



Inducing Translation Lexicon:
No Parallel Data (Haghighi et al., 2008)

- Orthographic (n-gram characters)
- Contextual features



Orthographic Features

#es	1.0
sta	1.0
do#	1.0

Context Features

mundo	17.0
economico	10.0
sociedad	6.0

Orthographic Features

#st	1.0
tat	1.0
te#	1.0

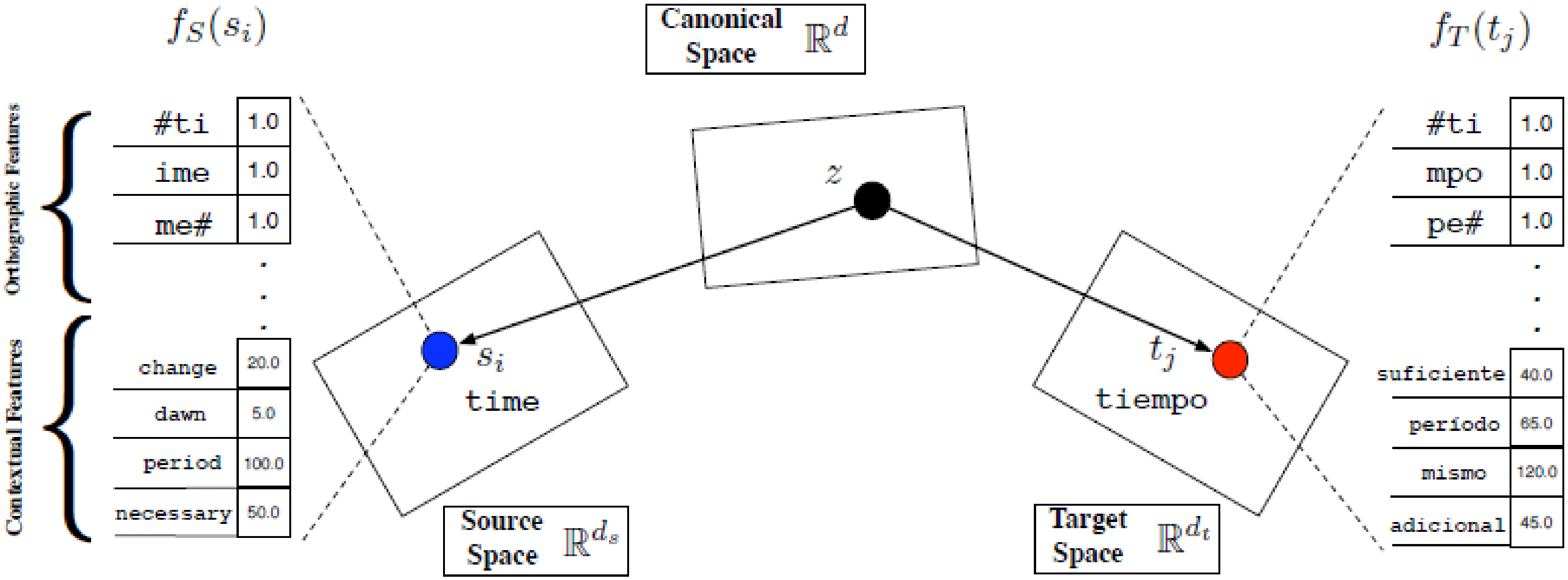
Context Features

world	17.0
economic	5.0
society	10.0

Inducing Translation Lexicon: No Parallel Data (Haghighi et al., 2008)

Observed words in source and target spaces are projected to a common latent space

Canonical Correlation Analysis (CCA)


$f_S(s_i)$
Orthographic Features
Contextual Features
Canonical Space $\mathbb{R}^d$
z
s_i
time
Source Space $\mathbb{R}^{d_s}$
t_j
tiempo
Target Space $\mathbb{R}^{d_t}$
$f_T(t_j)$

#ti	1.0
ime	1.0
me#	1.0
⋮	⋮
change	20.0
dawn	5.0
period	100.0
necessary	50.0

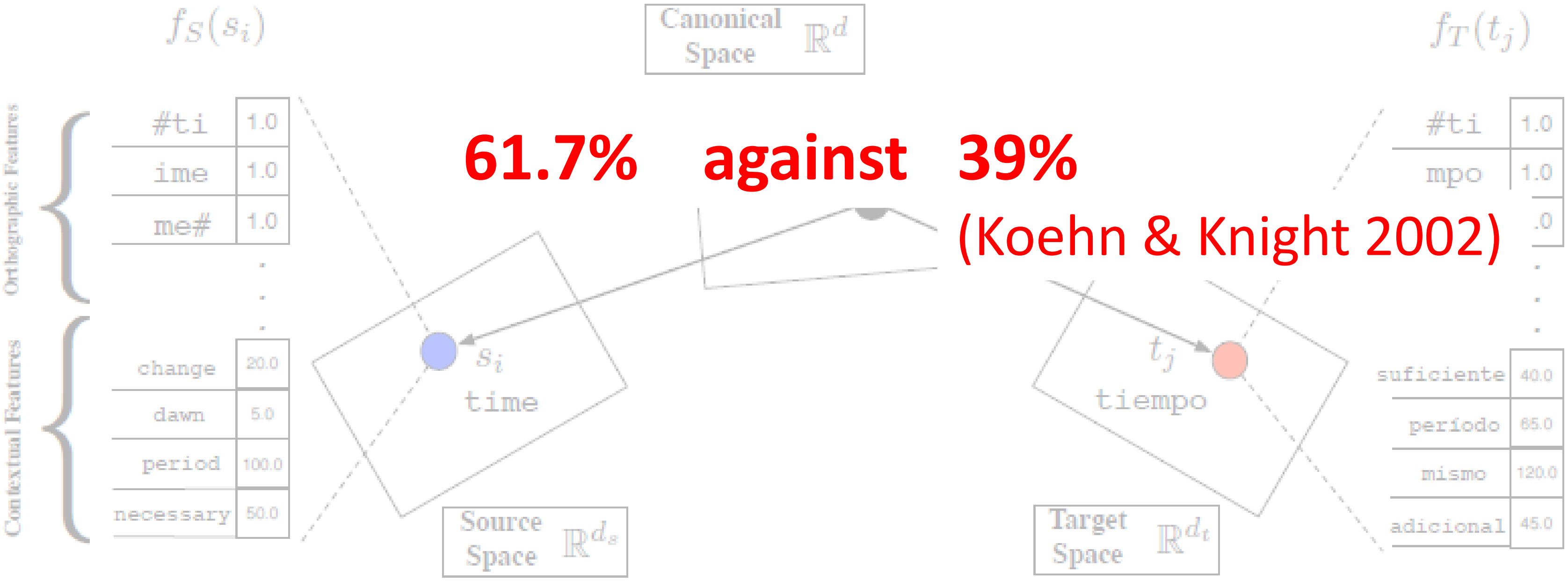
#ti	1.0
mpo	1.0
pe#	1.0
⋮	⋮
suficiente	40.0
período	65.0
mismo	120.0
adicional	45.0

(Haghighi et al., 2008)

Inducing Translation Lexicon: No Parallel Data (Haghighi et al., 2008)

Observed words in source and target spaces are projected to a common latent space

Canonical Correlation Analysis (CCA)



61.7% against 39%

(Koehn & Knight 2002)

(Haghighi et al., 2008)

Methods

- Inducing Translation Lexicon
  - Limited Parallel Data
  - Monolingual Data

- Translation System
  - Word-based
  - Phrase-based

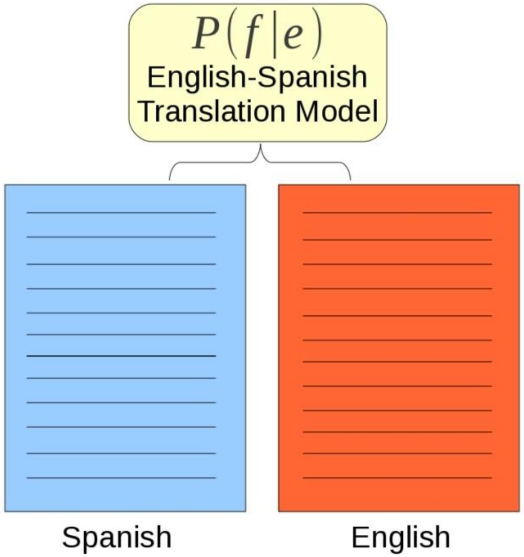
Translation System:

Word-based: MT as decipherment (Ravi and Knight 2011)

f: source sentence **e: target sentence**

Train parameters to maximize probability of **observed sentence pairs (e,f)**:

$$\underset{\theta}{argmax}\, P_{\theta}(e,f) \simeq \underset{\theta}{argmax} \prod_{e,f} P_{\theta}(f|e)$$



$P(f|e)$

English-Spanish Translation Model

Spanish

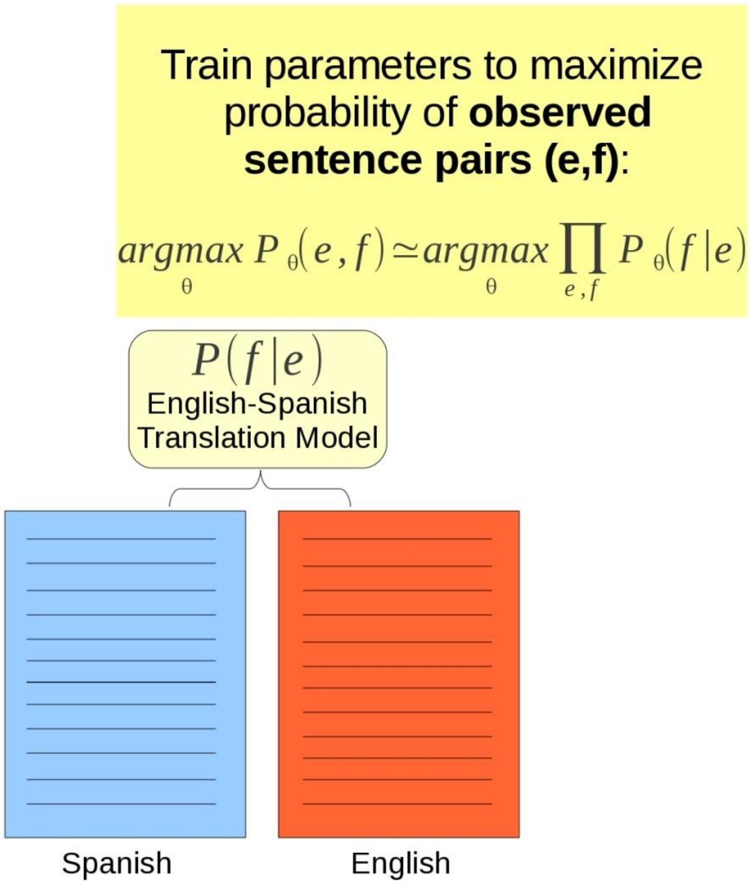
English

Training with Parallel data

Translation System:

Word-based: MT as decipherment (Ravi and Knight 2011)

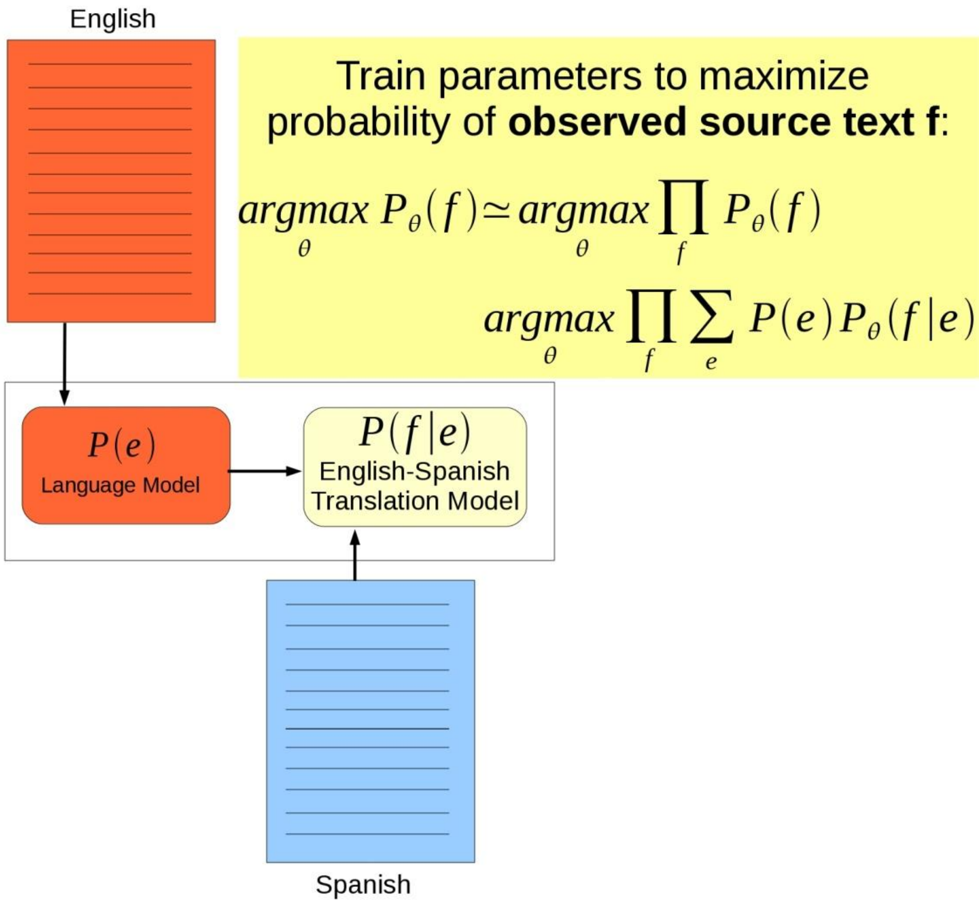
f: source sentence **e: target sentence**



Train parameters to maximize probability of **observed sentence pairs (e,f)**:

$$\underset{\theta}{argmax}\, P_{\theta}(e,f) \simeq \underset{\theta}{argmax} \prod_{e,f} P_{\theta}(f|e)$$

Training with Parallel data



Train parameters to maximize probability of **observed source text f**:

$$\underset{\theta}{argmax}\, P_{\theta}(f) \simeq \underset{\theta}{argmax} \prod_{f} P_{\theta}(f)$$

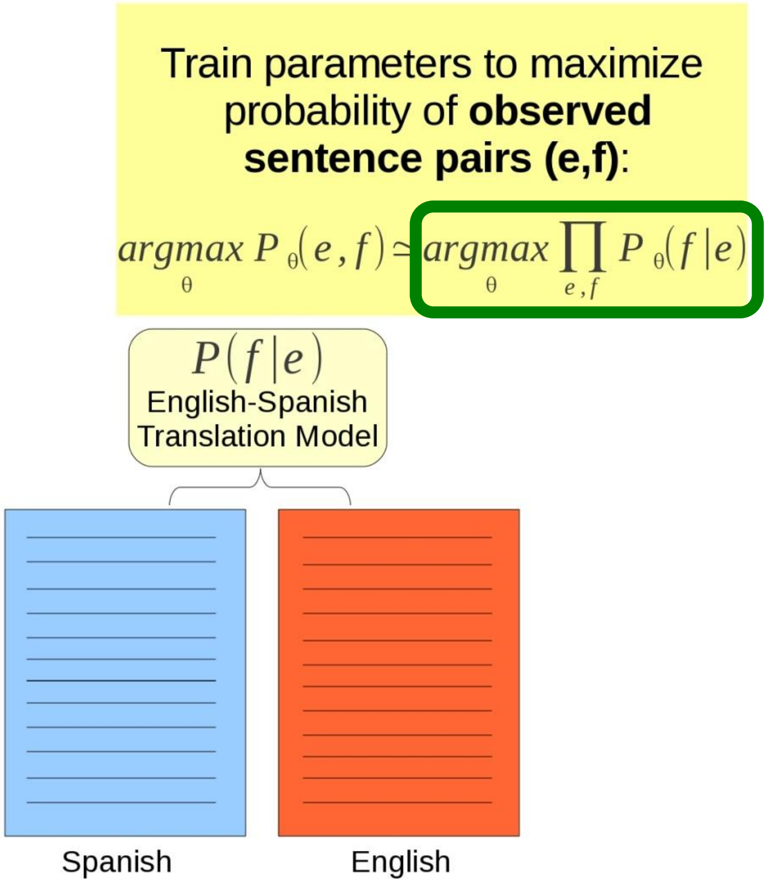
$$\underset{\theta}{argmax} \prod_{f} \sum_{e} P(e) P_{\theta}(f|e)$$

Training without Parallel data

Translation System:

Word-based: MT as decipherment (Ravi and Knight 2011)

f: source sentence **e: target sentence**



Train parameters to maximize probability of **observed sentence pairs (e,f)**:

$$\underset{\theta}{argmax}\, P_{\theta}(e,f) \simeq \boxed{\underset{\theta}{argmax} \prod_{e,f} P_{\theta}(f|e)}$$

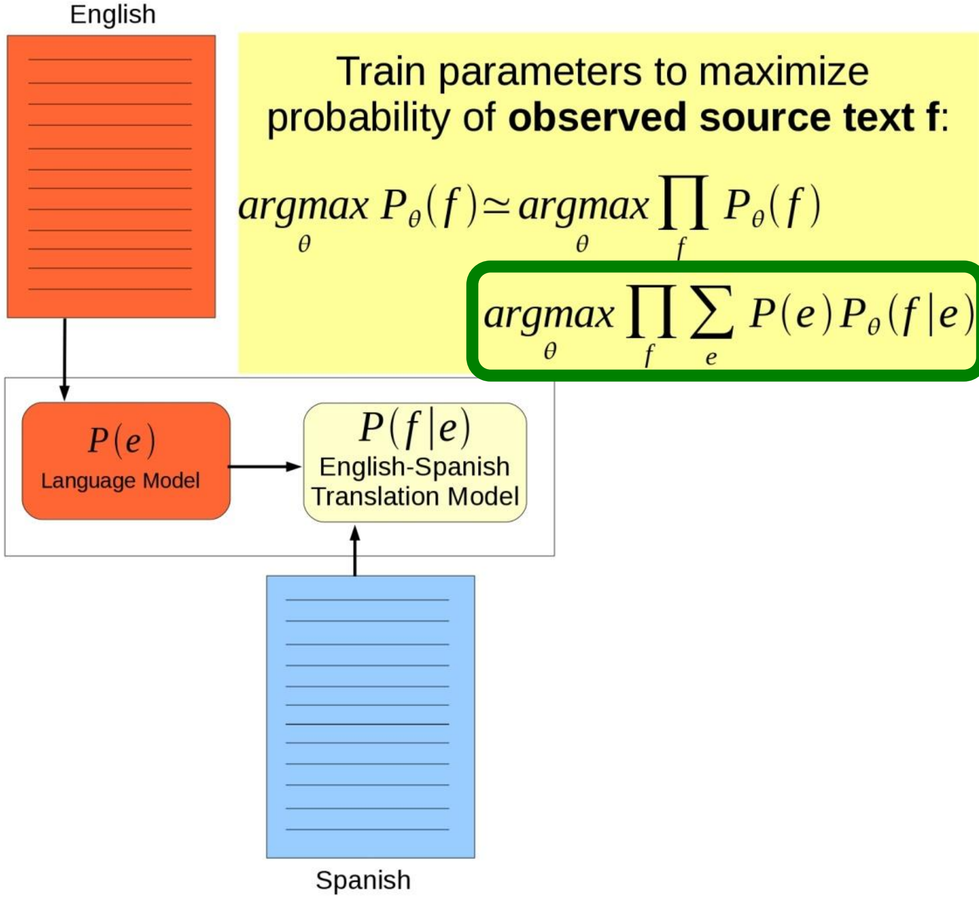
$P(f|e)$

English-Spanish Translation Model

Spanish

English

Training with Parallel data



English

Train parameters to maximize probability of **observed source text f**:

$$\underset{\theta}{argmax}\, P_{\theta}(f) \simeq \underset{\theta}{argmax} \prod_{f} P_{\theta}(f)$$

$$\boxed{\underset{\theta}{argmax} \prod_{f} \sum_{e} P(e) P_{\theta}(f|e)}$$

$P(e)$

Language Model

$P(f|e)$

English-Spanish Translation Model

Spanish

Training without Parallel data

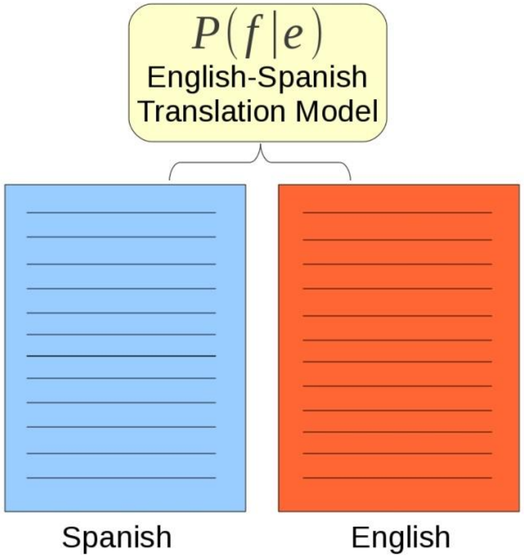
Translation System:

Word-based: MT as decipherment (Ravi and Knight 2011)

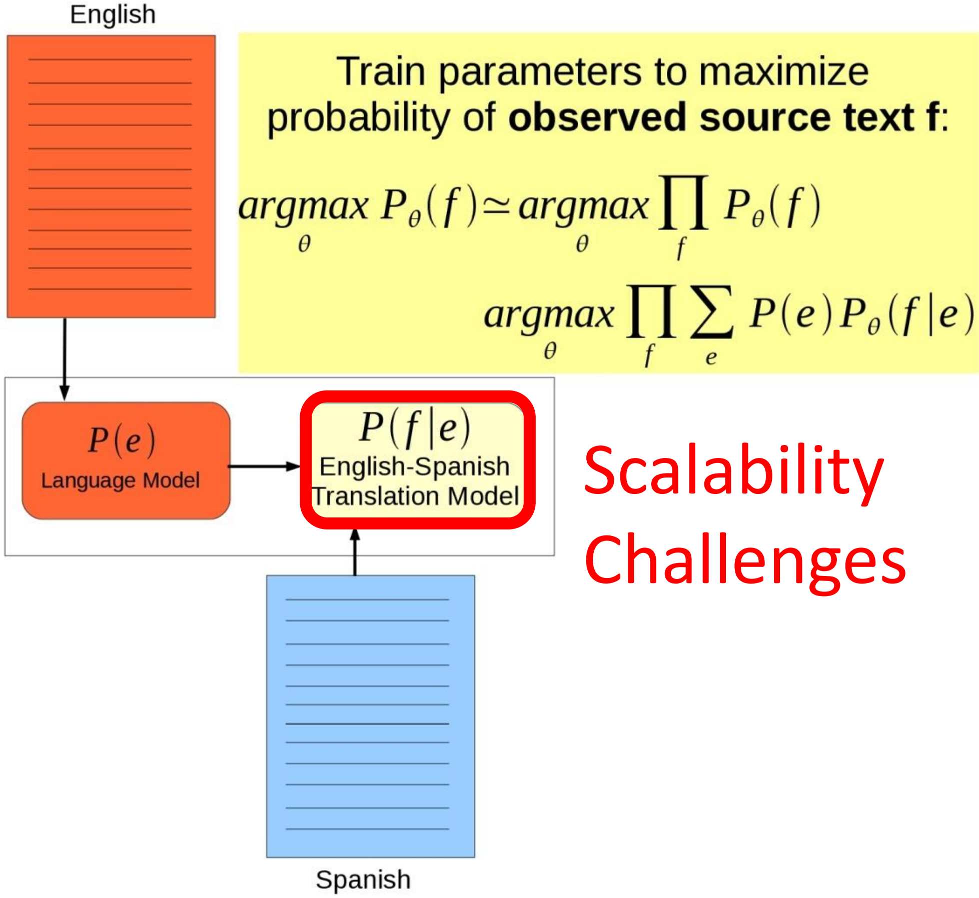
f: source sentence **e: target sentence**

Train parameters to maximize probability of **observed sentence pairs (e,f)**:

$$\underset{\theta}{argmax}\, P_\theta(e,f) \simeq \underset{\theta}{argmax} \prod_{e,f} P_\theta(f|e)$$



Training with Parallel data



Train parameters to maximize probability of **observed source text f**:

$$\underset{\theta}{argmax}\, P_\theta(f) \simeq \underset{\theta}{argmax} \prod_f P_\theta(f)$$

$$\underset{\theta}{argmax} \prod_f \sum_e P(e) P_\theta(f|e)$$

Scalability Challenges

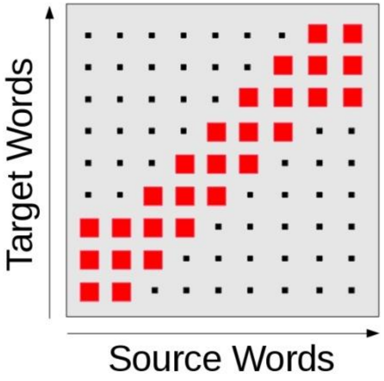
Training without Parallel data

Translation System: Word-based

- Restricting the translation model (Nuhn et al., 2012)
  - Determining a set of active translations
  - Estimating the probabilities of active translations

Translation System: Word-based

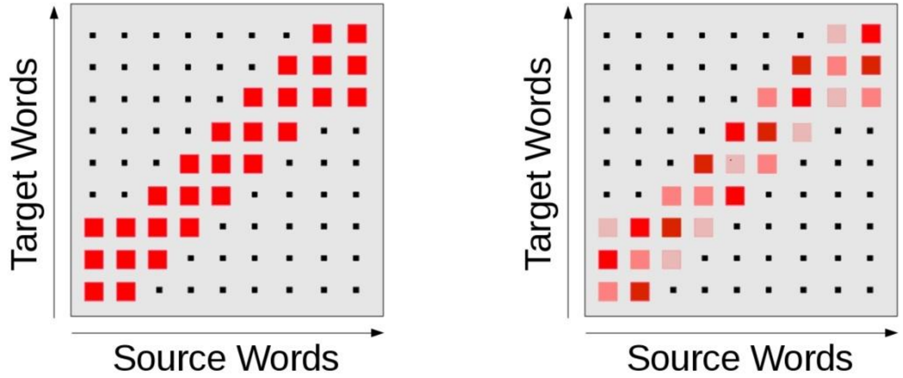
- Restricting the translation model (Nuhn et al., 2012)
  - Determining a set of active translations
  - Estimating the probabilities of active translations



Initialization using word frequency ranks

Translation System: Word-based

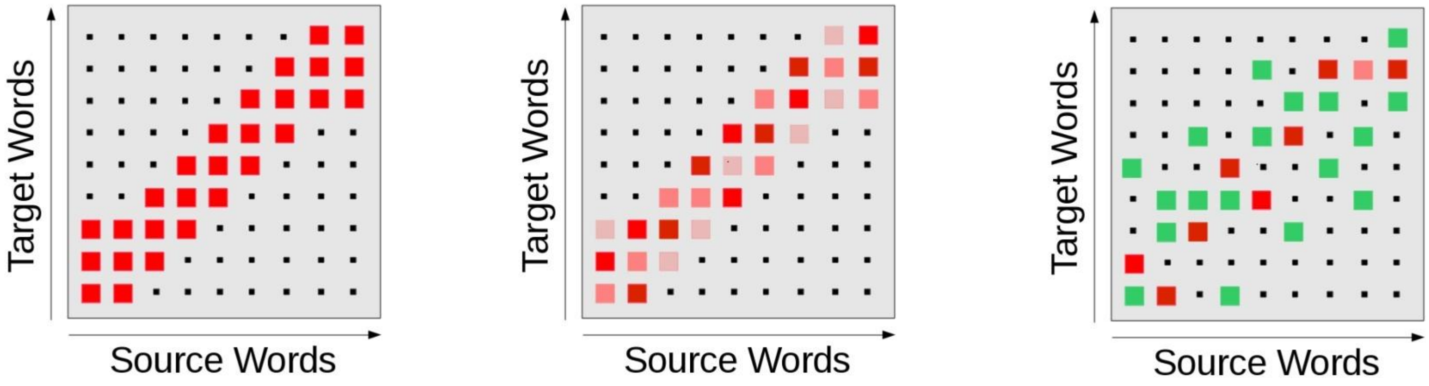
- Restricting the translation model (Nuhn et al., 2012)
  - Determining a set of active translations
  - Estimating the probabilities of active translations



  - EM training (Ravi and Knight 2011) with limited iterations (20-30)

Translation System: Word-based

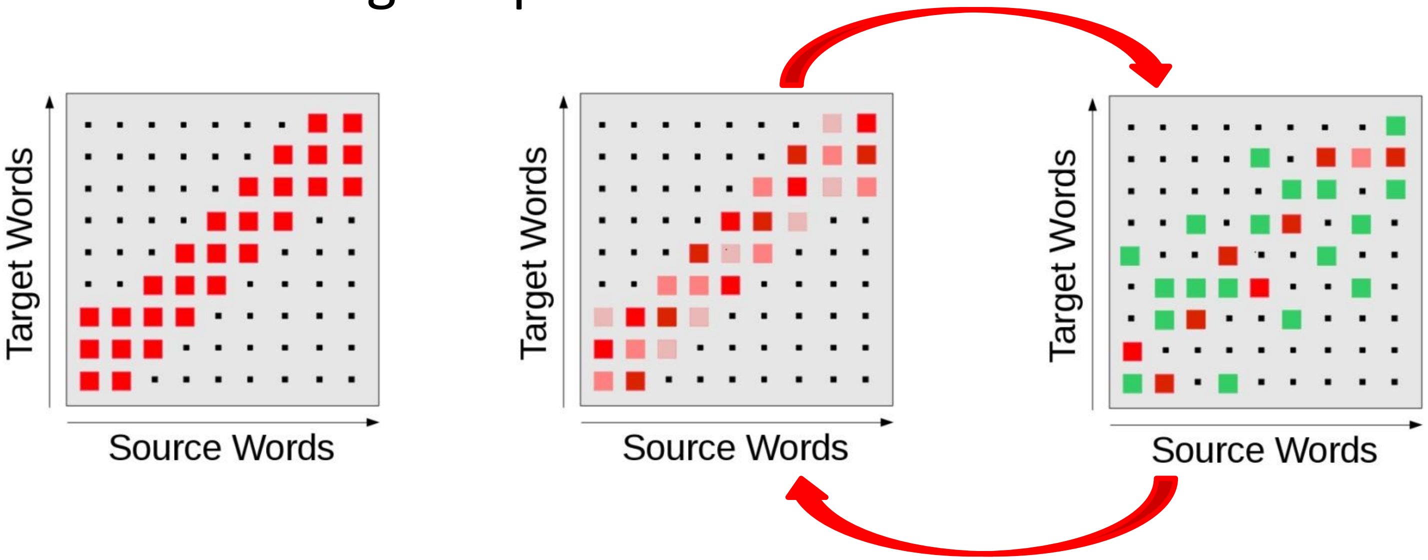
- Restricting the translation model (Nuhn et al., 2012)
  - Determining a set of active translations
  - Estimating the probabilities of active translations



Inducing active translations using results of EM

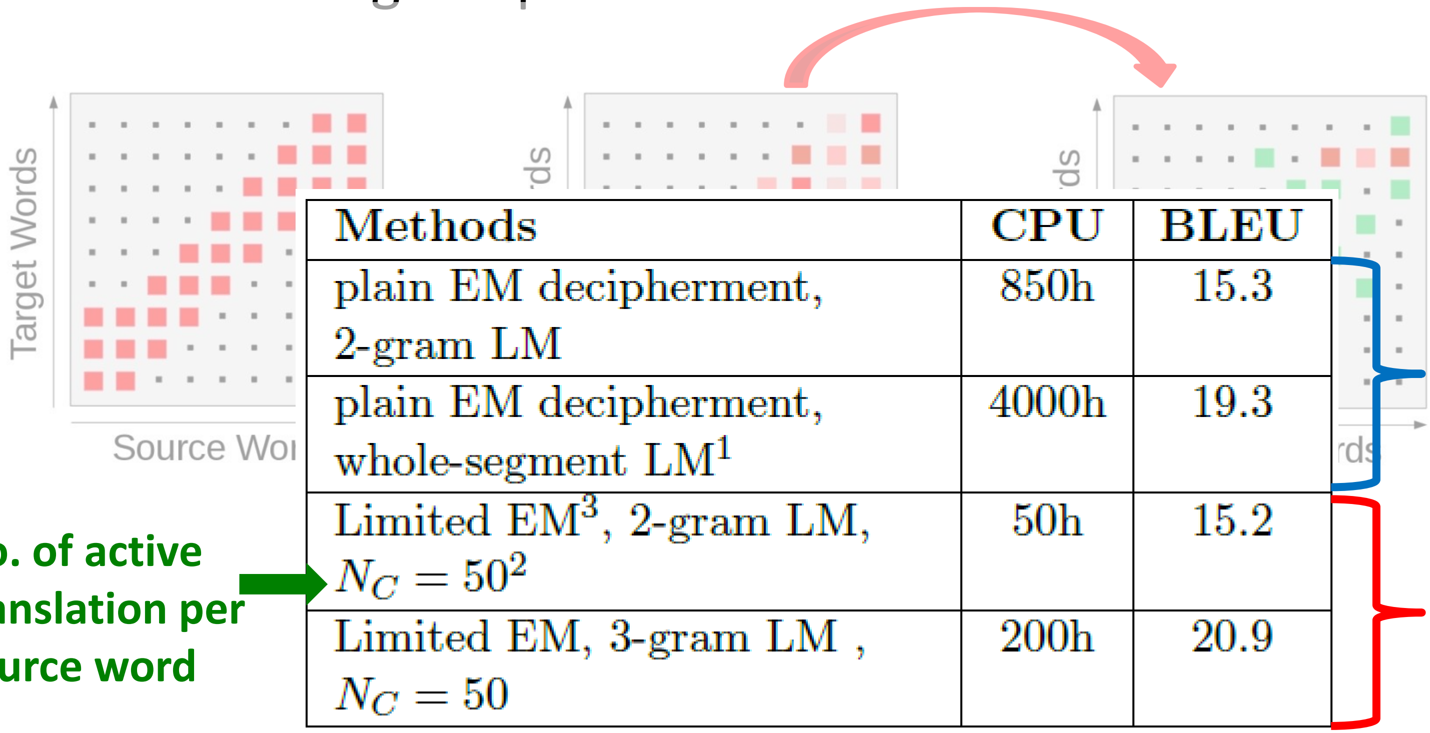
Translation System: Word-based

- Restricting the translation model (Nuhn et al., 2012)
  - Determining a set of active translations
  - Estimating the probabilities of active translations



Translation System: Word-based

- Restricting the translation model (Nuhn et al., 2012)
  - Determining a set of active translations
  - Estimating the probabilities of active translations


Target Words
Source Wor
rds
ds
rds

Methods	CPU	BLEU
plain EM decipherment, 2-gram LM	850h	15.3
plain EM decipherment, whole-segment LM[1]	4000h	19.3
Limited EM[3], 2-gram LM, $N_C = 50^2$	50h	15.2
Limited EM, 3-gram LM , $N_C = 50$	200h	20.9

No. of active translation per source word

Ravi & Knight 2011

Nuhn et al., 2012

Translation System

Limited Parallel Data

Phrase-based Translation System: Limited Parallel Data(Klementiev et al.,2012)

- Parameters in phrase-based translation systems
  - Large amounts of parallel data for estimating parameters

Phrase-based Translation System: Limited Parallel Data(Klementiev et al.,2012)

- Parameters in phrase-based translation systems
  - Large amounts of parallel data for estimating parameters
- Extract features from monolingual data
  - Extend the idea of translation lexicon induction to phrases (using seed lexicon)
  - Reordering model

Phrase-based Translation System: Limited Parallel Data(Klementiev et al.,2012)

- Parameters in phrase-based translation systems
  - Large amounts of paralle g

Bilingual Features

21.87 BLEU score

from m

Monolingual &

Bilingual Features

22.92 BLEU score

  - Extend the idea of translation lexicon induction to phrases (using seed lexicon)
  - Reordering model

Conclusion

- Translation lexicon induction:
  - Depend on context and orthographic similarities
    - Does not work on historically unrelated language pairs
  - Scalability: Just applied on small word sets(1000 or so)

- Translation Systems:
  - New research direction
  - Scalability: applicable to limited vocabularies and data sets
  - Low translation quality