USC University of Southern California

Joint Word Alignment and Decipherment Improves Machine Translation

Qing Dou, Ashish Vaswani, and Kevin Knight

10/26/2014

Outline

- What is Decipherment
- Motivation
- Contributions
- Joint Word Alignment and Decipherment
- Deciphering Malagasy
- Conclusions

What is Decipherment?

- Letter Substitution Cipher

plaintext

d e c i p h e r m e n t _ i s _ t h e _ a n
a l y s i s _ o f _ d o c u m e n t s _ w r i
t t e n _ i n _ a n c i e n t _ l a n g u a g
e s

What is Decipherment?

- Letter Substitution Cipher

plaintext

d e c i p h e r m e n t _ i s _ t h e _ a n
a l y s i s _ o f _ d o c u m e n t s _ w r i
t t e n _ i n _ a n c i e n t _ l a n g u a g
e s

Encryption

ciphertext

05 13 19 25 12 14 13 04 02 13 11 16 15
25 22 15 16 14 13 15 17 11 17 08 03 22
25 22 15 09 20 15 05 09 19 07 02 13 11
22 15 06 04 25 16 16 13 11 15 25 11 15
17 11 19 25 13 11 16 15 08 17 11 03 07
17 03 13 22

What is Decipherment?

- Letter Substitution Cipher

plaintext

d e c i p h e r m e n t _ i s _ t h e _ a n
a l y s i s _ o f _ d o c u m e n t s _ w r i
t t e n _ i n _ a n c i e n t _ l a n g u a g
e s

Encryption

ciphertext

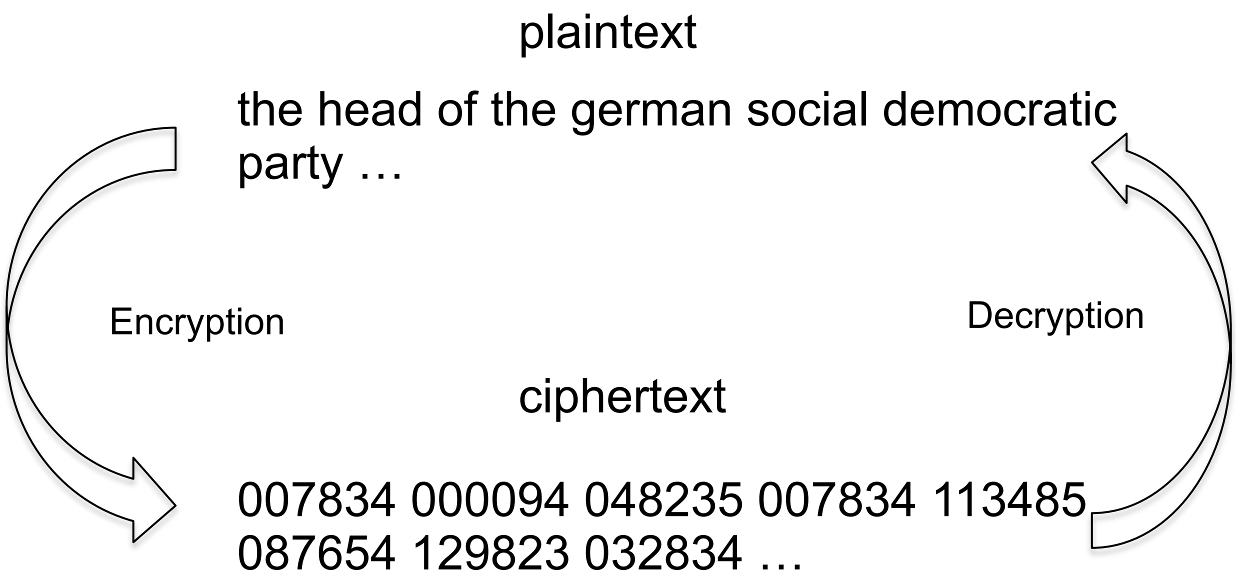
05 13 19 25 12 14 13 04 02 13 11 16 15
25 22 15 16 14 13 15 17 11 17 08 03 22
25 22 15 09 20 15 05 09 19 07 02 13 11
22 15 06 04 25 16 16 13 11 15 25 11 15
17 11 19 25 13 11 16 15 08 17 11 03 07
17 03 13 22

Decipherment

a	**17**	**o**	**09**
b	01	p	12
c	19	q	23
d	05	r	04
[illegible]	[illegible]	s	22
[illegible]	20	t	16
[illegible]	10	u	07
h	14	v	24
l	25	w	06
j	18	x	26
k	21	y	03
l	08	z	27
m	02	_	15
n	11		

Substitution Cipher and Translation

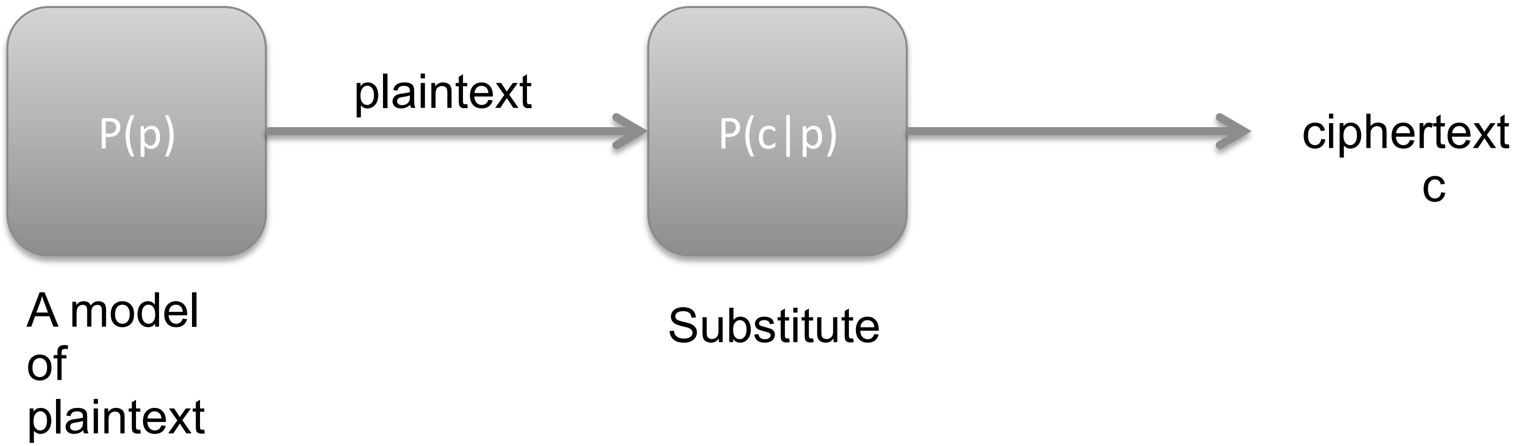
- Word Substitution Cipher



- Word substitutions also take place in translation

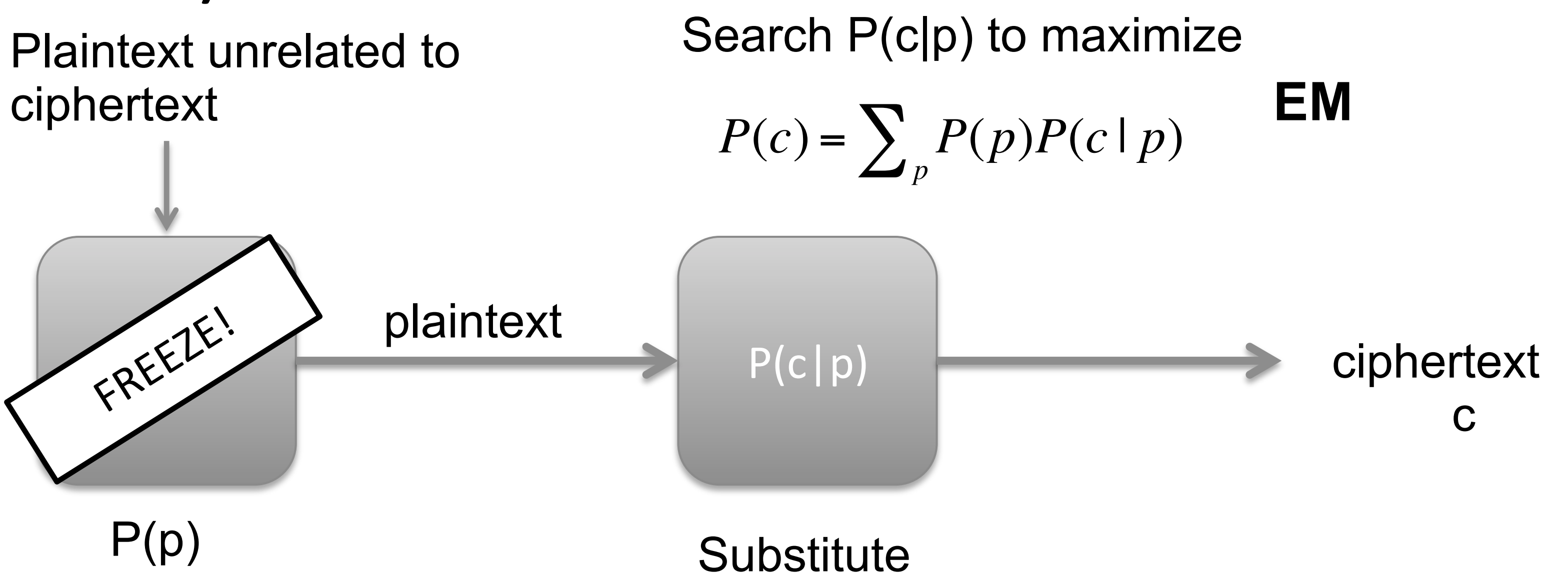
Automatic Decipherment

- A Noisy Channel Model Approach (Knight et al. 2006)



Automatic Decipherment

- A Noisy Channel Model Approach (Knight et al. 2006)



Automatic Decipherment

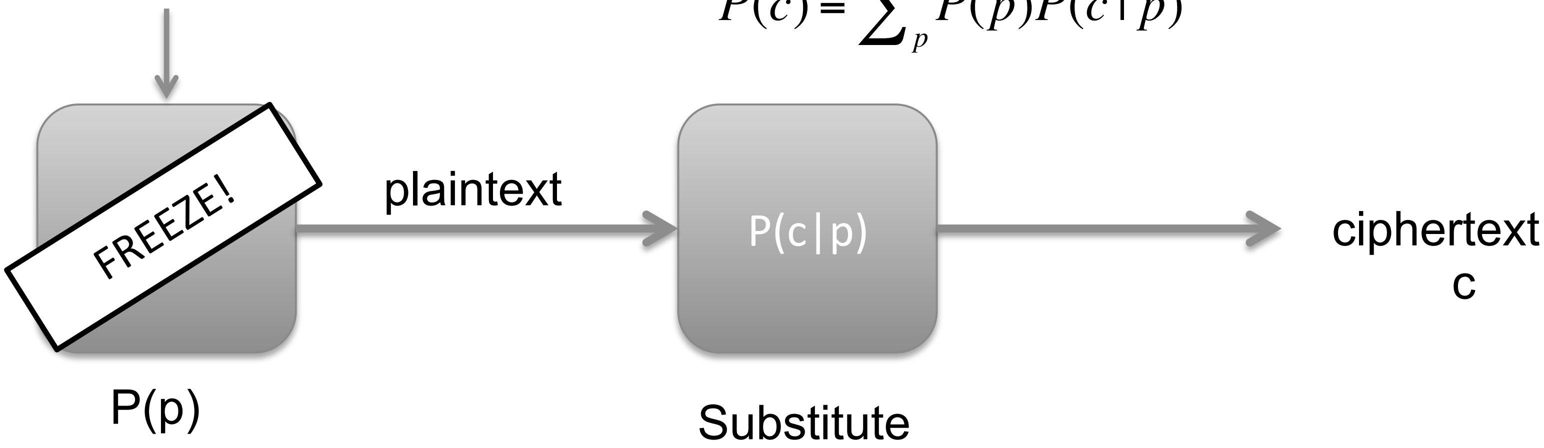
- A Noisy Channel Model Approach (Knight et al. 2006)

Search P(c|p) to maximize

$$P(c) = \sum_p P(p)P(c \mid p)$$

EM

Plaintext unrelated to ciphertext



- Time Complexity: $O(N \cdot V^2 \cdot R)$

(Forward-backward)

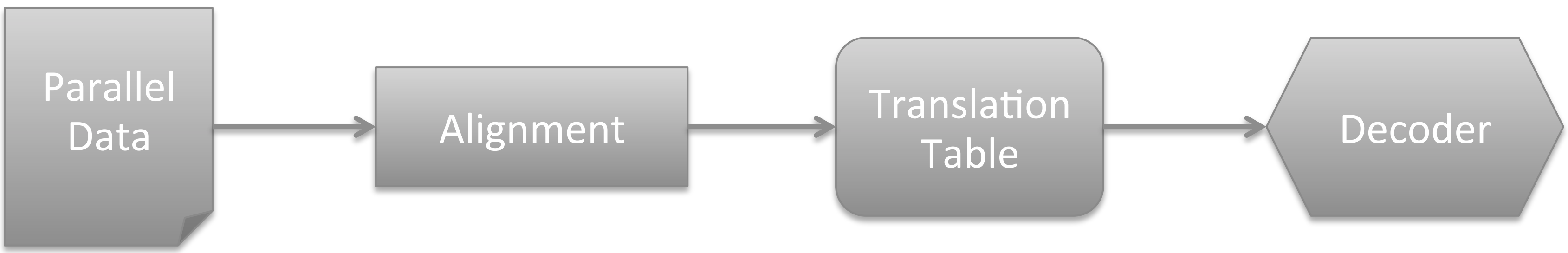
N: Ciphertext length
V: Vocabulary
R: EM iteration

Outline

- What is Decipherment
- Motivation
- Contributions
- Joint Word Alignment and Decipherment
- Deciphering Malagasy
- Conclusions

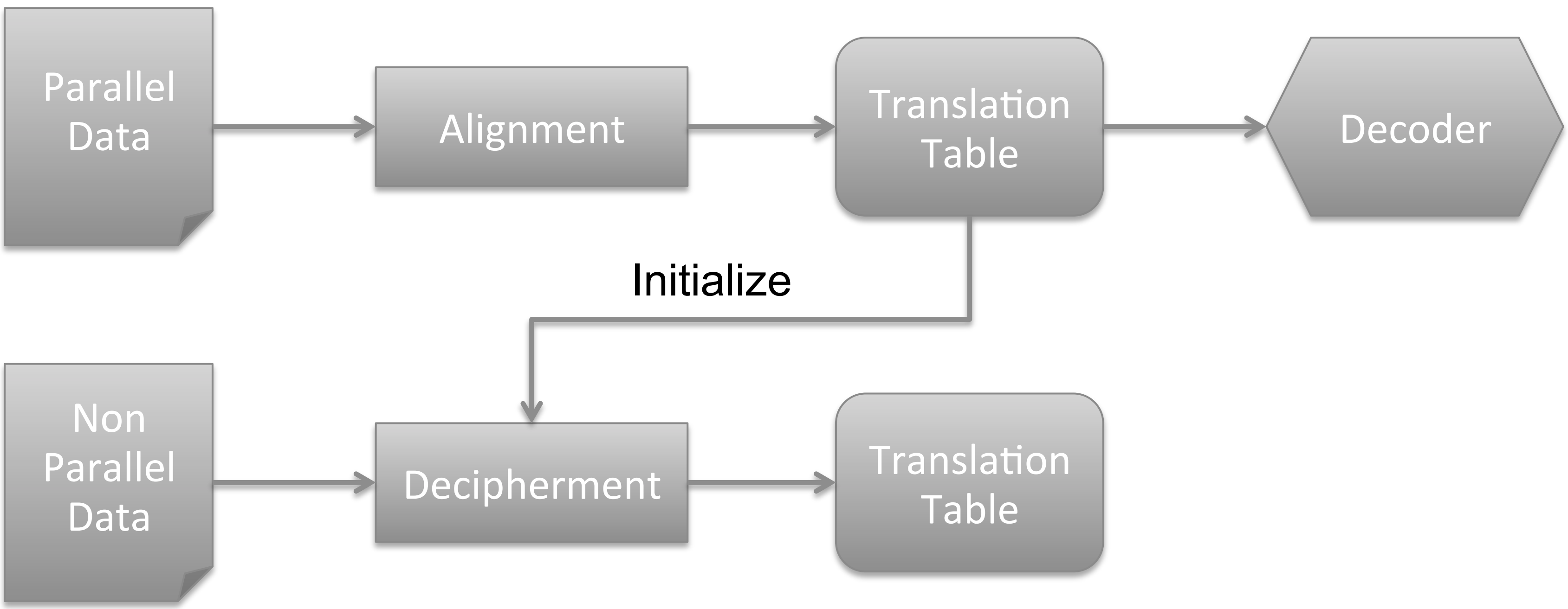
Motivation

- Decipherment improves machine translation (Dou and Knight 2013)



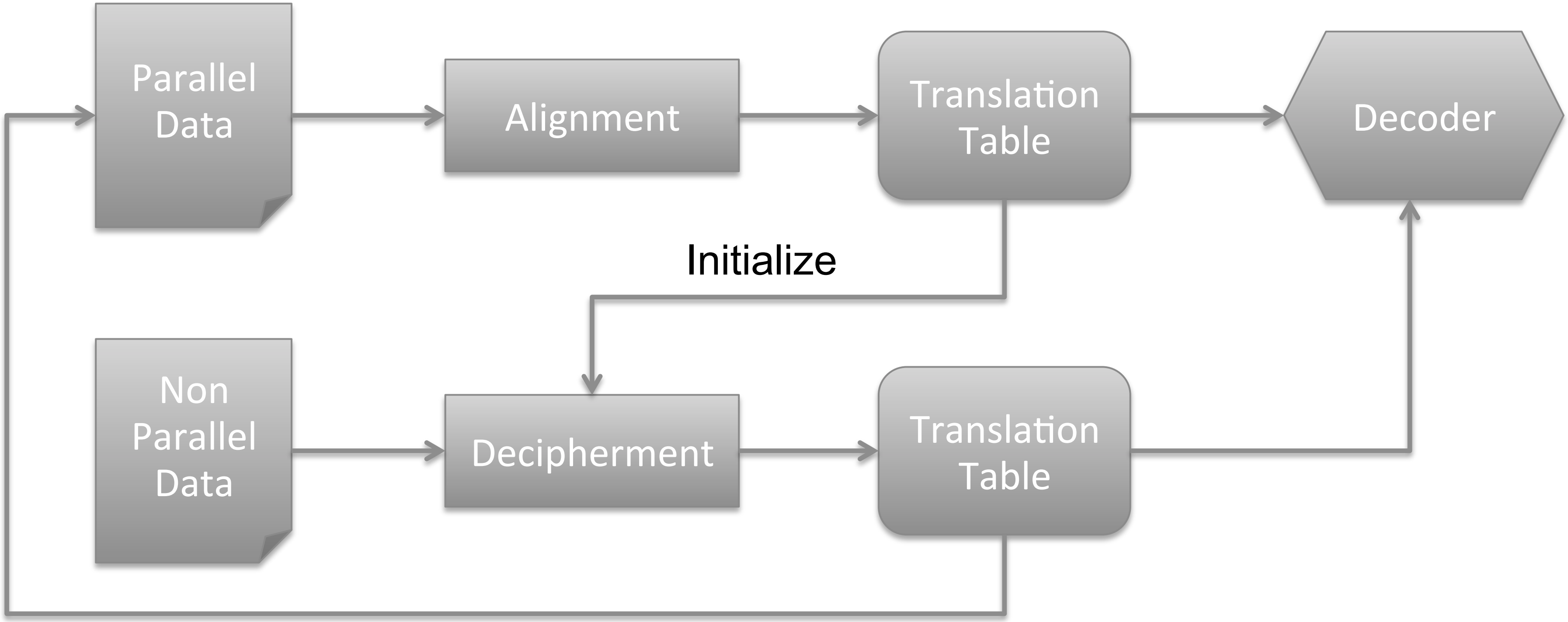
Motivation

- Decipherment improves machine translation (Dou and Knight 2013)



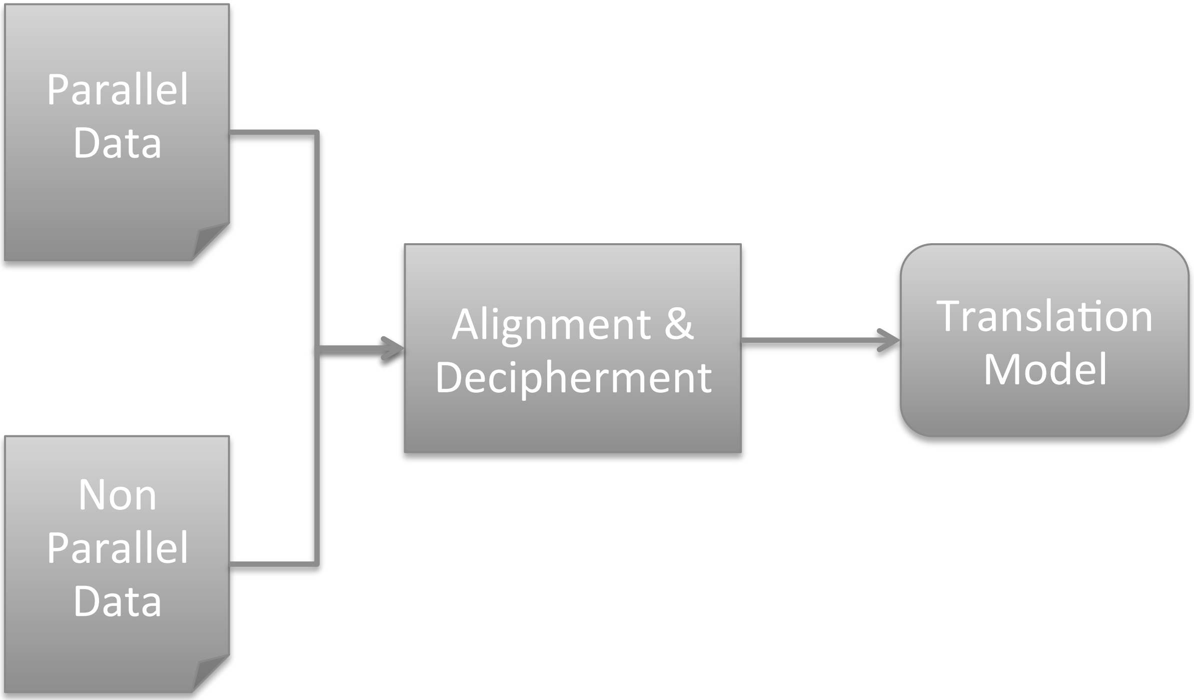
Motivation

- Decipherment improves machine translation (Dou and Knight 2013)



Motivation

- Joint Alignment and Decipherment ?



Contributions

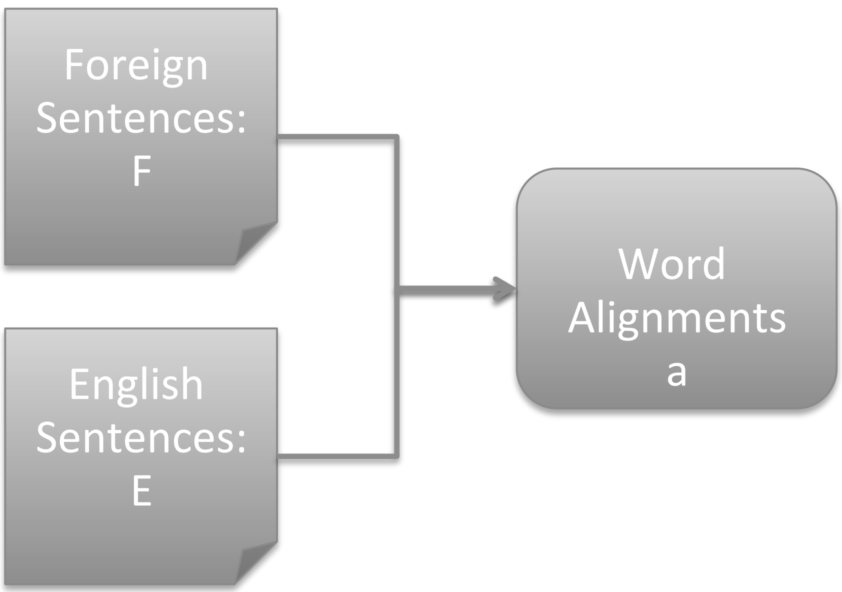
- Proposed a new framework to perform joint word alignment and decipherment
- The joint framework improves both word alignment and machine translation significantly
- Released Malagasy treebank and 15.3 million word Malagasy news data

Outline

- What is Decipherment
- Motivation
- Contributions
- Joint Word Alignment and Decipherment
- Deciphering Malagasy
- Conclusions

Word Alignment

- Word Alignment Model and Objective



Objective:

$$P(F \mid E) = \sum_{a} \prod_{j=1}^{J} \underbrace{d(a)}_{\text{distortion probabilities}} \cdot \underbrace{t(f_j \mid e_{a_j})}_{\text{translation probabilities}}$$

Decipherment

- Decipherment Model and Objective

(Dependency based Decipherment Dou and Knight 2013)

Foreign Dependency Bigrams: F_{mono} → English Decipherment e

Objective:

$$P(F_{mono}) = \sum_{e} P(e_1 e_2) \prod_{j=1}^{2} t(f_j \mid e_j)$$

LM Probabilities (fixed)

translation probabilities

LM: Dependency Language Model
(Created from dependency trees)

A New Objective

Word Alignment Objective:

$$P(F \mid E) = \sum_{a} \prod_{j=1}^{J} d(a) \cdot t(f_j \mid e_{a_j})$$

Decipherment Objective:

$$P(F_{mono}) = \sum_{e} P(e_1 e_2) \prod_{j=1}^{2} t(f_j \mid e_j)$$

Shared Parameters

$$t(f \mid e)$$

A New Objective

Word Alignment Objective:

$$P(F \mid E) = \sum_{a} \prod_{j=1}^{J} d(a) \cdot t(f_j \mid e_{a_j})$$

Decipherment Objective:

$$P(F_{mono}) = \sum_{e} P(e_1 e_2) \prod_{j=1}^{2} t(f_j \mid e_j)$$

Shared Parameters

$$t(f \mid e)$$

New Objective:

$$P(JOINT) = P(F \mid E) + \alpha P(F_{mono})$$

Learning Algorithm

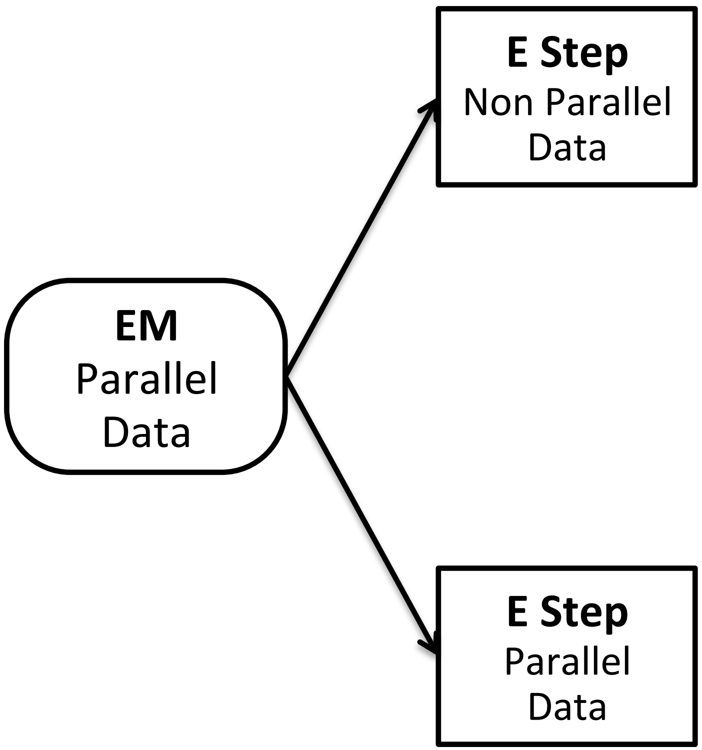
- EM

5 iterations of EM on
Parallel text only

EM
Parallel
Data

Learning Algorithm

- EM



Collect expected counts for:

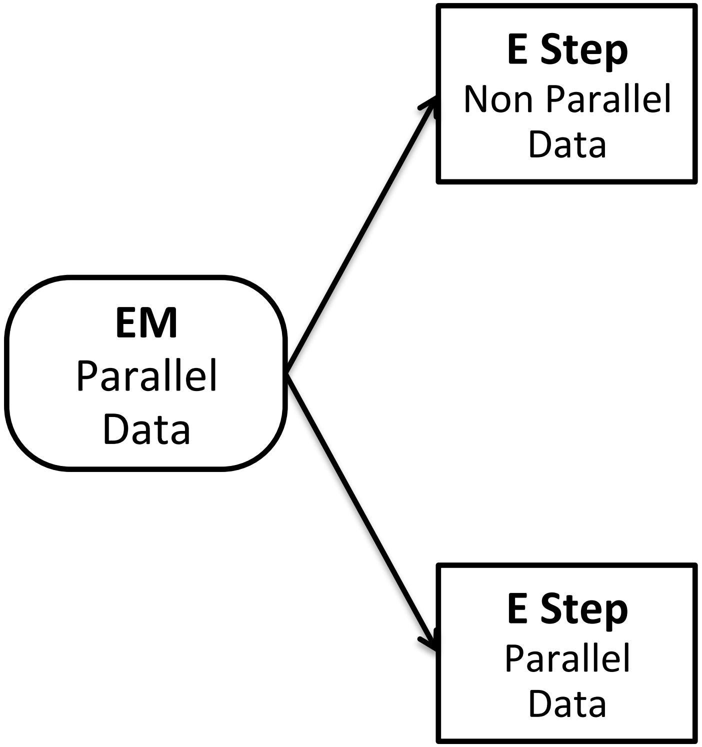
$t(f \mid e)$

Collect expected counts for:

$t(f \mid e) \quad d(a)$

Learning Algorithm

- EM



Collect expected counts for:

$t(f \mid e)$

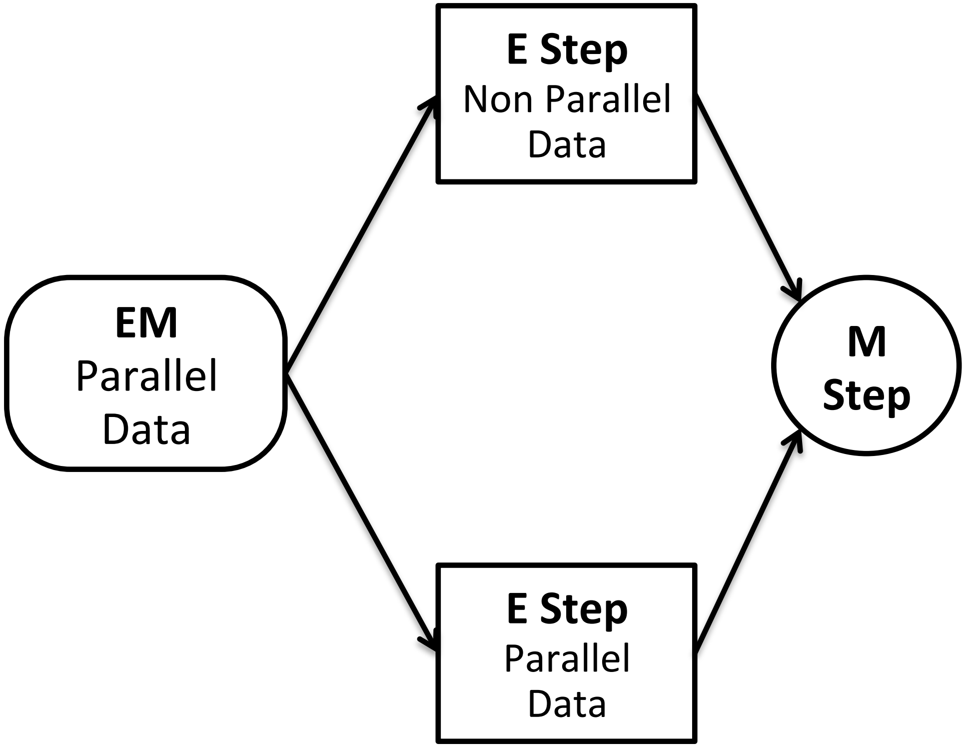
Collect expected counts for:

$t(f \mid e) \quad d(a)$

Sum up expected counts

Learning Algorithm

- EM

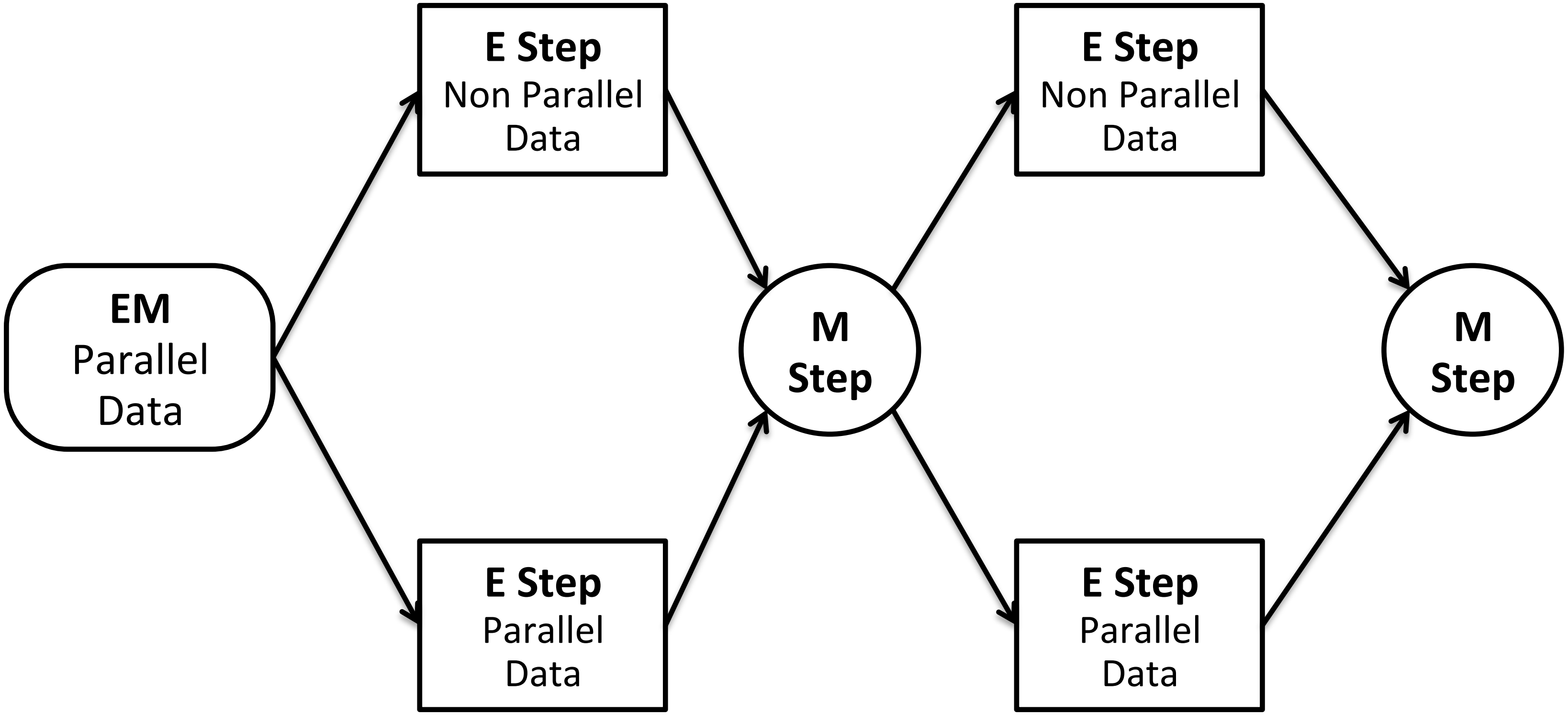


Update parameters

$t(f \mid e) \qquad d(a)$

Learning Algorithm

- EM



E Step

- On Parallel Data

  (Brown et al. 1993, Vogel and Ney 1996)

E Step

- On Parallel Data

  (Brown et al. 1993, Vogel and Ney 1996)

- On Non-parallel Data

  Time complexity: $O(N \cdot V^2 \cdot R)$

  V: Vocabulary size N: Ciphertext length

- Not Scalable when $V \sim 10^5, N \sim 10^7$

E Step

- On Non-parallel Data

Use samples to collect expected counts:

cipher bigram　　　　　　　　　　　　　　sample bigram

$$f_1 f_2 \xrightarrow[\text{Slice Sampling}]{\propto P(f_1 \mid e_1)\cdot P(f_2 \mid e_2)\cdot P(e_1 e_2)} e_1 e_2$$

Let N be total number of samples we draw

And $e_1 e_2$ be one of them:

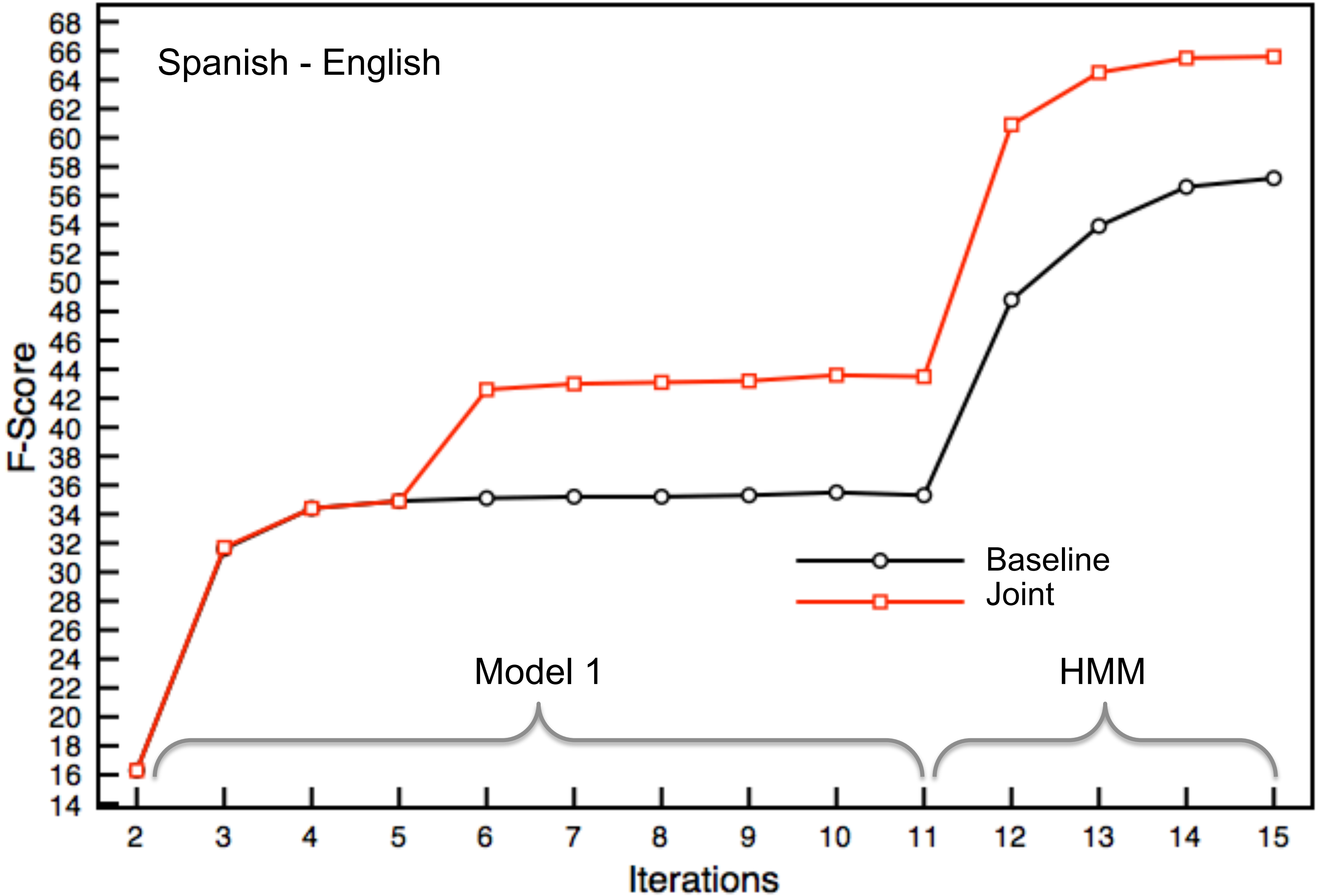
$$Expected\_Count(f_1, e_1) = Expected\_Count(f_2, e_2) = \frac{1}{N} \cdot count(f_1, f_2)$$

Word Alignment Experiment

- Data (Size in tokens)

	Spanish	**English**
Parallel	10.3k	9.9k
Non Parallel	80 million	400 million
TreeBank	0.4 million	1.0 million

Decipherment Improves Alignment



Outline

- What is Decipherment
- Motivation
- Contributions
- Joint Word Alignment and Decipherment
- Deciphering Malagasy
- Conclusions

The Malagasy Language

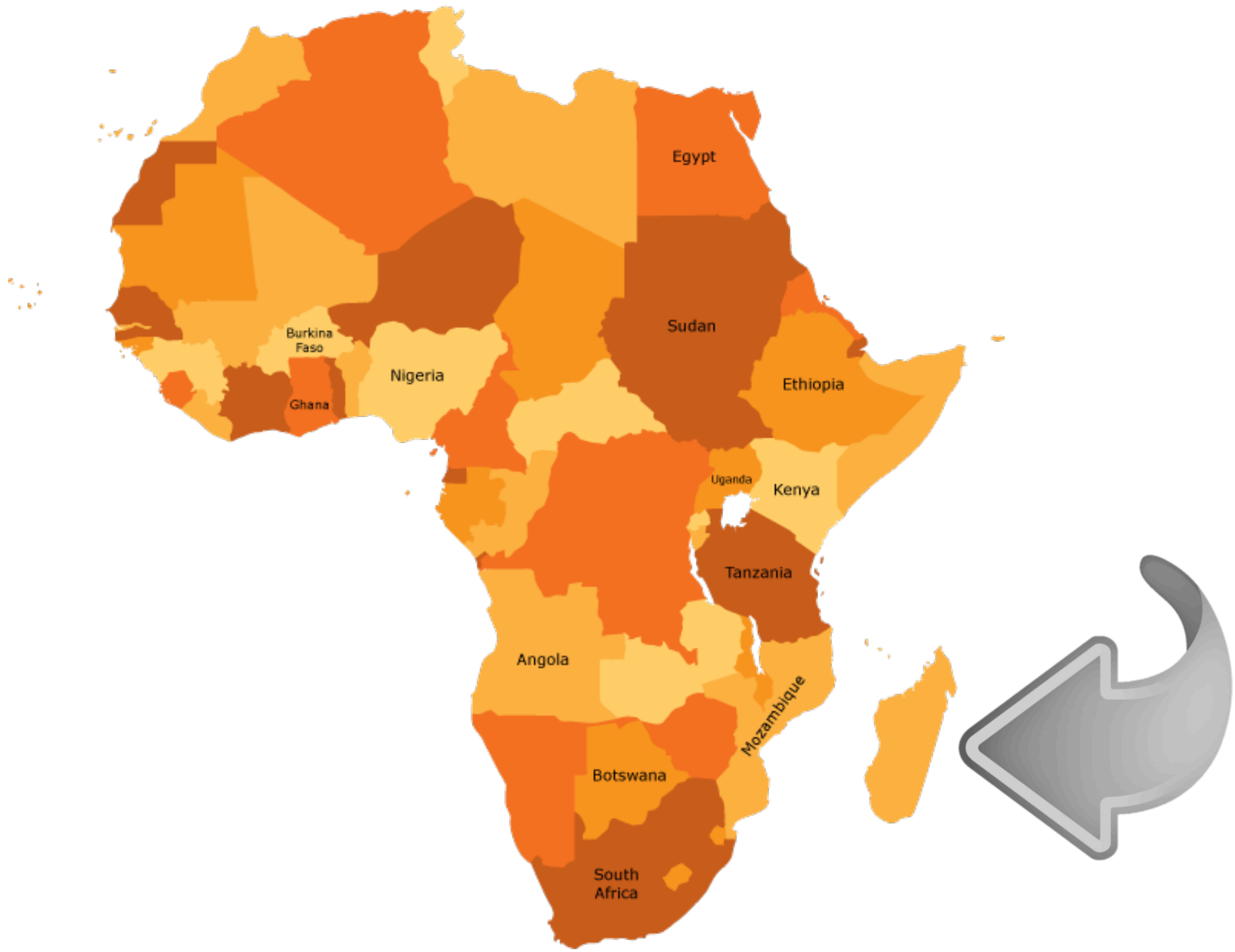
- Is official Language of Madagascar

The Malagasy Language

- Is official Language of Madagascar

The Malagasy Language

- Is official Language of Madagascar



The Malagasy Language

- Is official Language of Madagascar
- Although spoken in African, Malagasy has its root in southeast Asia.

The Malagasy Language

- Is official Language of Madagascar
- Although spoken in African, Malagasy has its root in southeast Asia.
- Has 18 million native speakers

The Malagasy Language

- Is official Language of Madagascar
- Although spoken in African, Malagasy has its root in southeast Asia.
- Has 18 million native speakers
- Is head initial with V-O-S word order. (English: S-V-O)

Malagasy Dependency Parser

- Data

Training	120 sentences, 20k tokens
Testing	48 sentences, 7k tokens

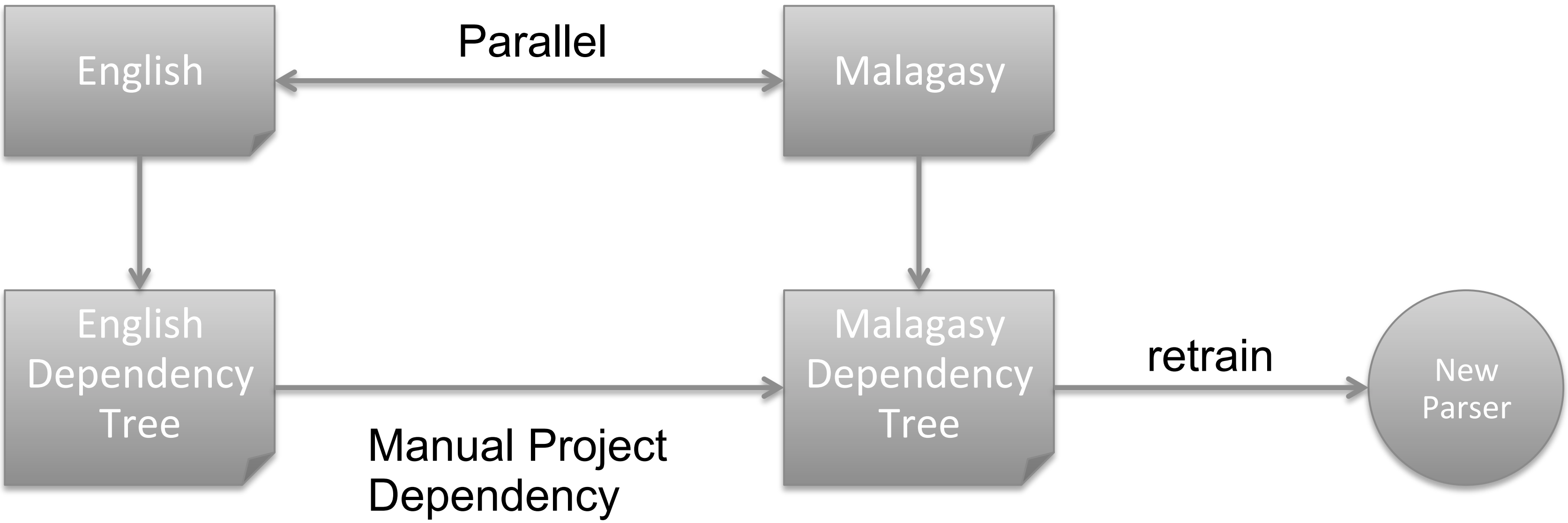
Spanish parser trained on 400k tokens

- Result on Malagasy

72.4 % directed attachment accuracy

Malagasy Dependency Parser

- More Training Data



- Result

  Improved to 80.0 % from 72.4%

Malagasy-English MT

- Data (In tokens)

		Malagasy	English
Parallel	Train (GV)	0.9 million	0.8 million
	Tune (GV)	22.2k	20.2k
	Test (GV)	23k	21k
	Test (Web)	2.2k	2.1k
Non Parallel	GigaWord	N/A	834 million
	Web	15.3 million	396 million

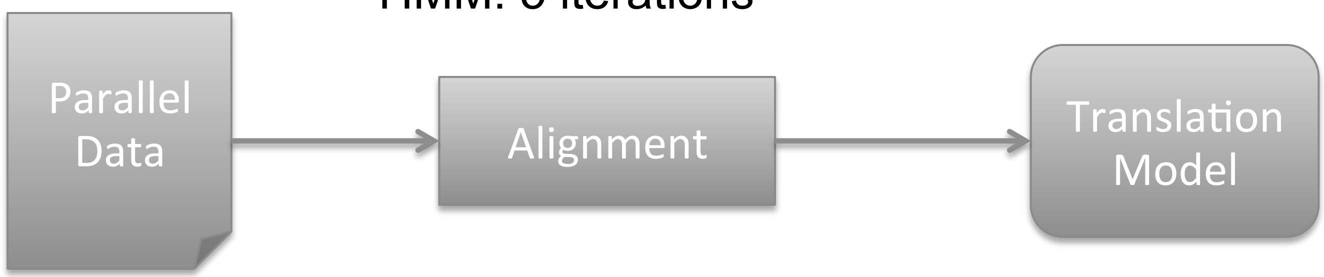
GV: Global Voices, multilingual international news website

Malagasy-English MT

- Baseline

Phrase-based MT system with Moeses

Model1: 10 iterations
HMM: 5 iterations

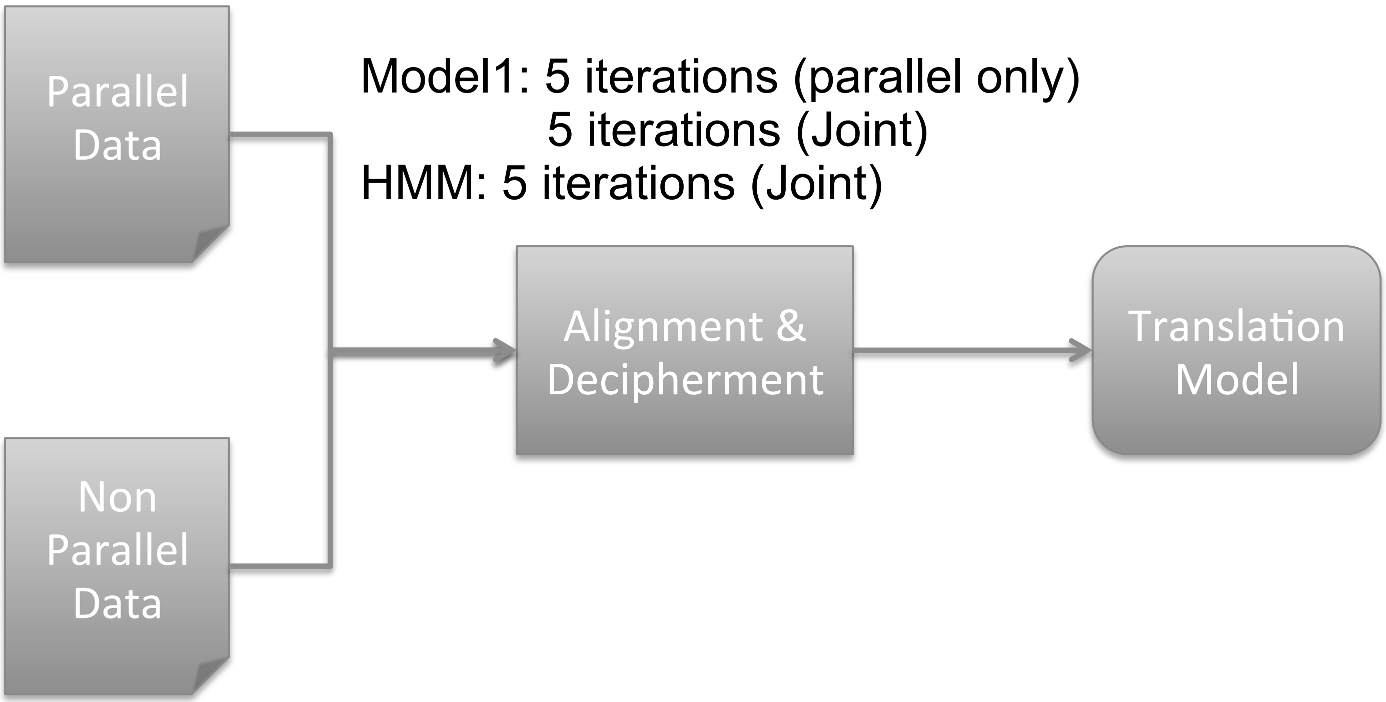


(Model 3 and Model 4 doesn't improve BLEU)

Align in 2 directions and used grow-diag-final to extract phrases

Malagasy-English MT

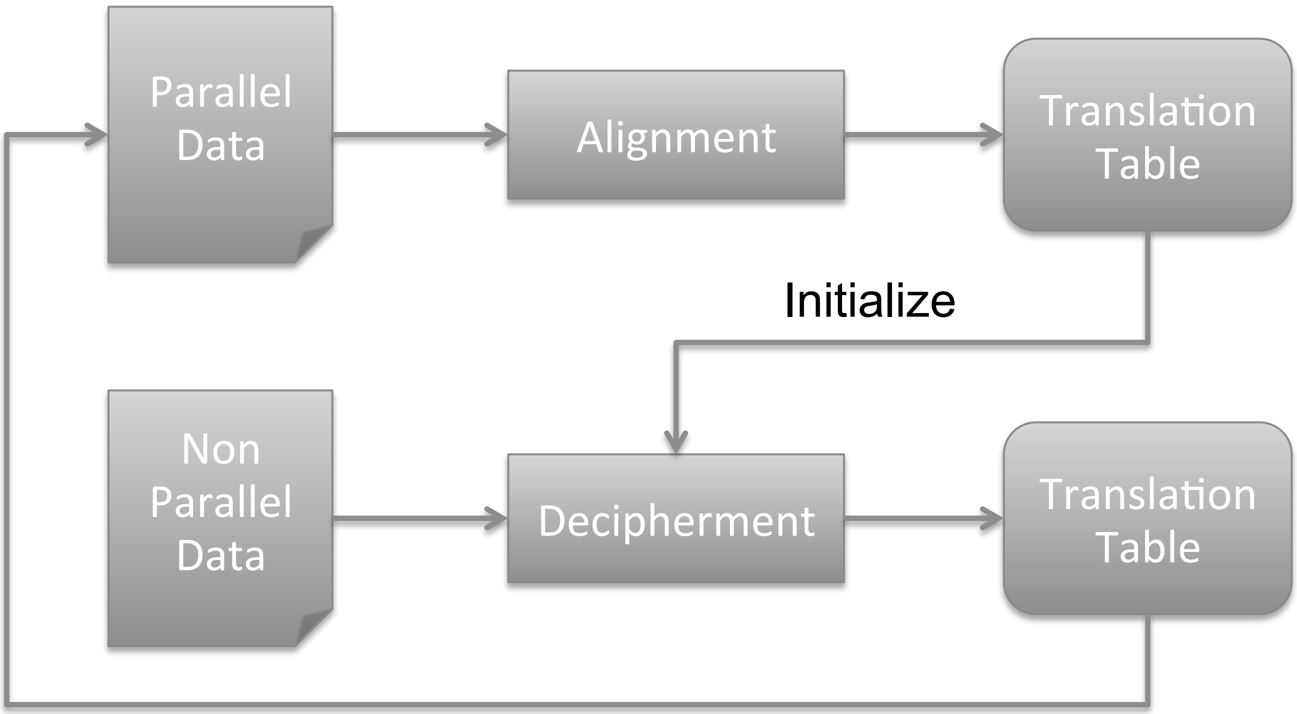
- Joint



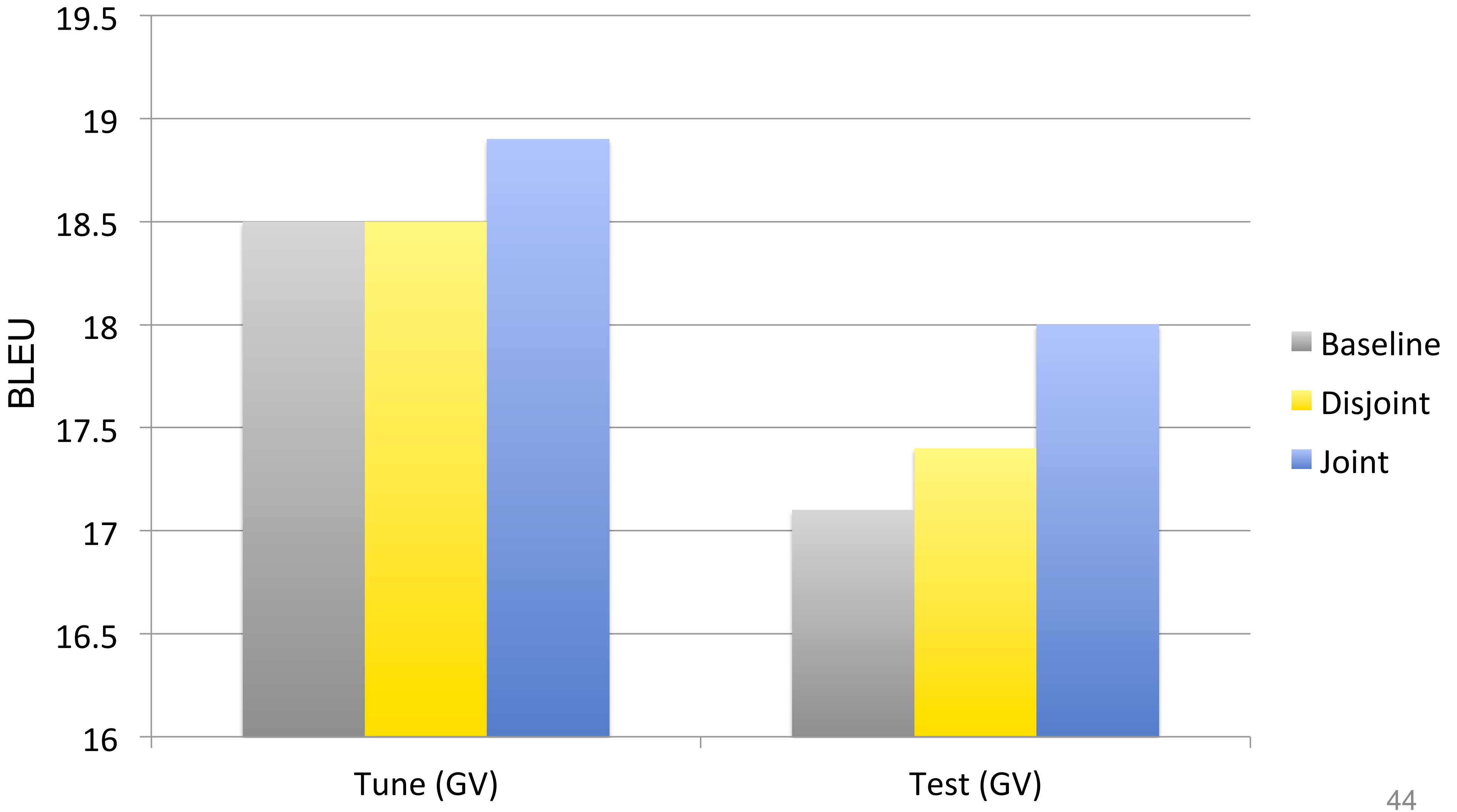
Align and extract phrases only on one direction P(English|Malagasy)

Malagasy-English MT

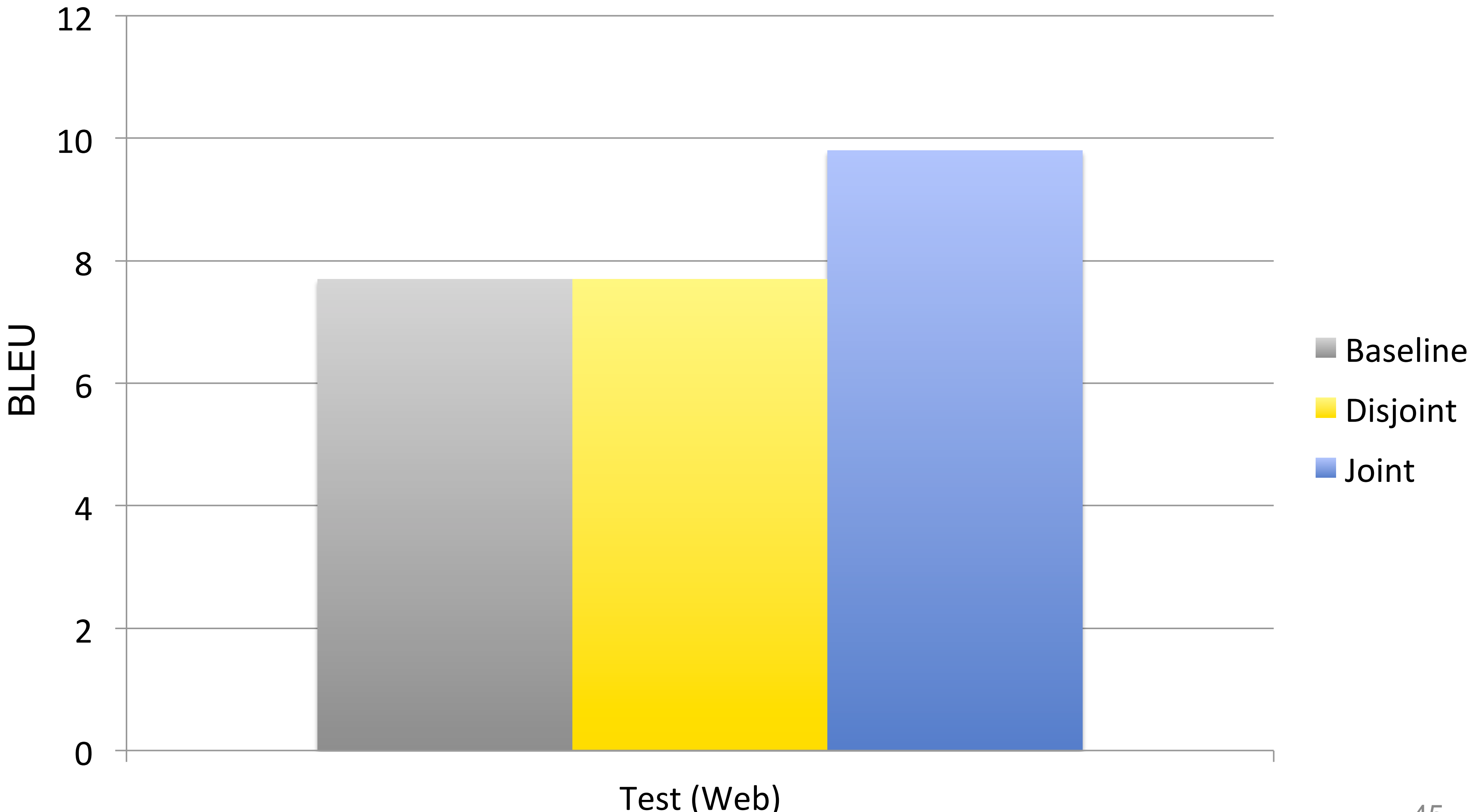
- Disjoint



Results on Global Voices



Results on Local News



Conclusion

- Proposed a framework for joint alignment and decipherment
- The joint process improves both alignment and machine translation quality
- Released a mini Malagasy treebank and 15m tokens news data

Thank You!