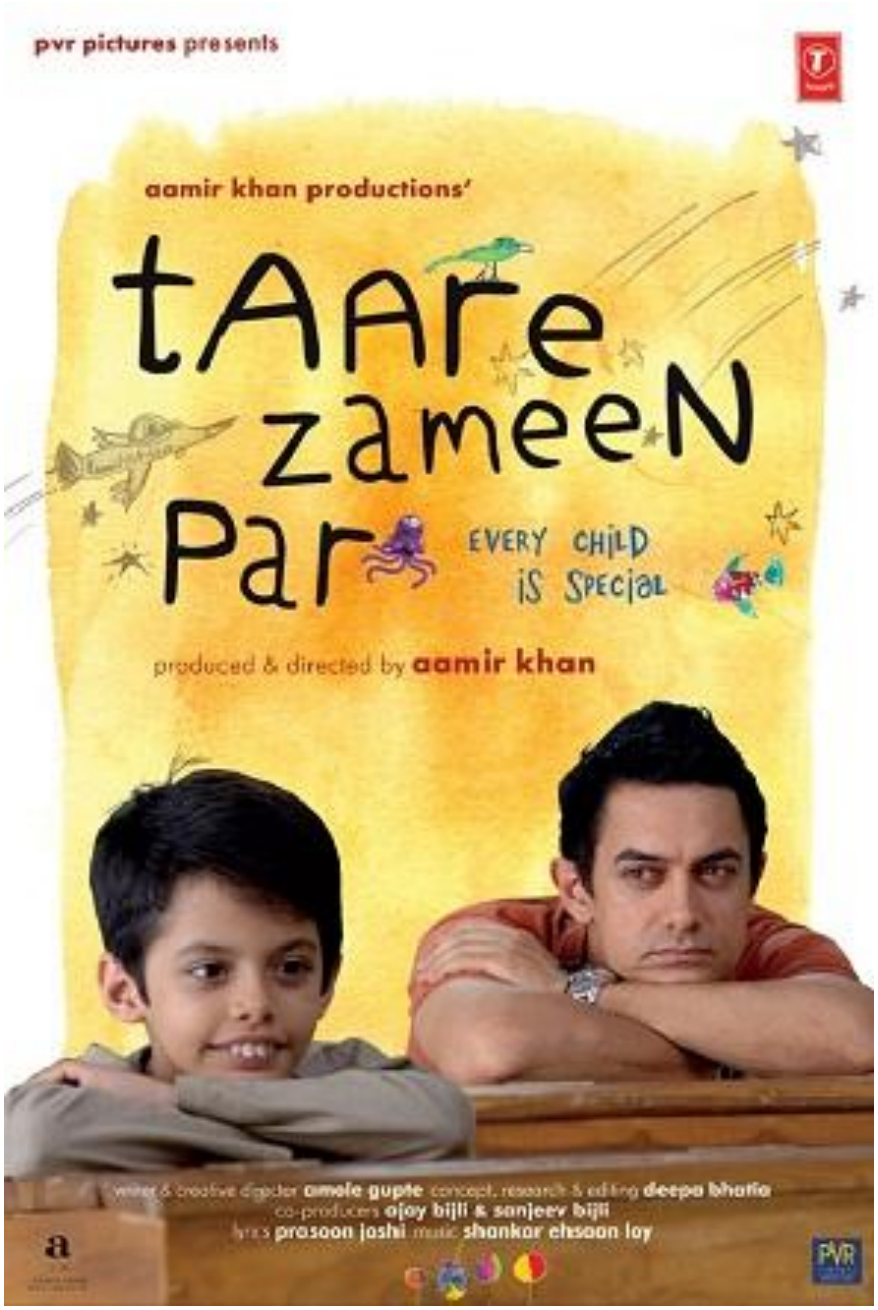
Using Multilingual Topic Models for Improved Alignment in English-Hindi MT

“GOOD TEACHERS KNOW HOW TO BRING OUT THE BEST IN STUDENTS”

– CHARLES KURALT

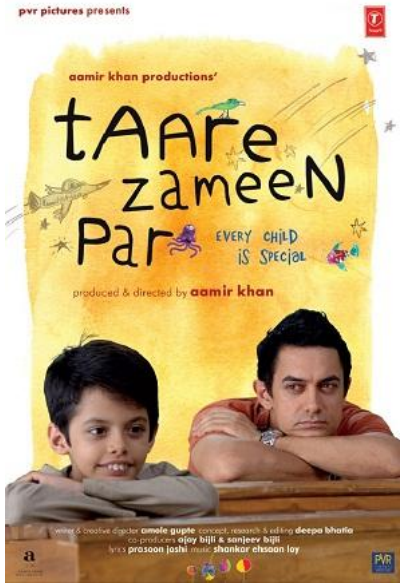
Acknowledgment:
CFILT, IIT Bombay
TCS Research Fellowship Program, TCS


pvr pictures presents
aamir khan productions'
tAAre zameeN Par
EVERY CHILD IS SPECIAL
produced & directed by aamir khan
writer & creative digester amole gupte concept, research & editing deepa bhatia
co-producers ajay bijli & sanjeev bijli
lyrics prasoon joshi music shankar ehsaan loy


DICTIONARY


A
अ

Using Multilingual Topic Models for Improved Alignment in English-Hindi MT



Diptesh Kanojia[1] **Aditya Joshi[1;2;3]**

Pushpak Bhattacharyya[1]

Mark James Carman[2]

[1]IIT Bombay, India, [2]Monash University, Australia

[3]IITB-Monash Research Academy, India

diptesh@cse.iitb.ac.in, adityaj@cse.iitb.ac.in

pb@cse.iitb.ac.in, mark.carman@monash.edu

What is the paper about?

- Dictionary is often appended to a parallel corpus ~~to improve alignment~~ Our paper
- A good dictionary is labour-intensive. Linguistic Data-driven

  How good is a coarse dictionary to improve alignment for MT?

- An approach may sound linguistically absurd but can still work

Outline

- Introduction & Dataset
- Multilingual Topic Models
- Multilingual Topics to Pseudo-parallel data
- Experimentation & Results
- Conclusion & Ongoing work

Introduction

- The right alignment is crucial
- We append pseudo-parallel data to a parallel corpus
- Our pseudo-parallel data is derived from a coarse dictionary obtained from a multilingual topic model

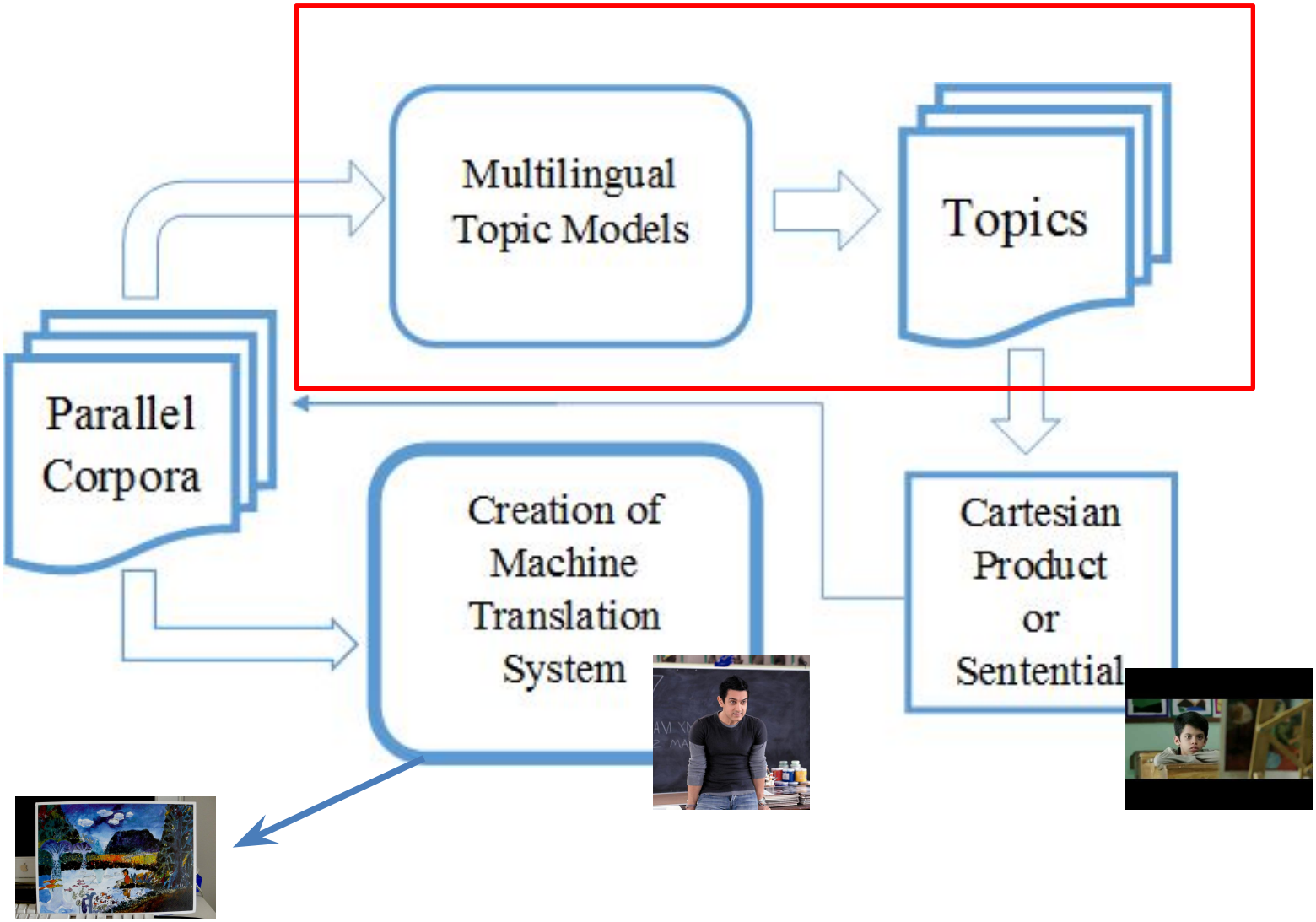
Dataset

- English-Hindi MT
- 25000 parallel English-Hindi sentences
- Health and tourism domain

Outline

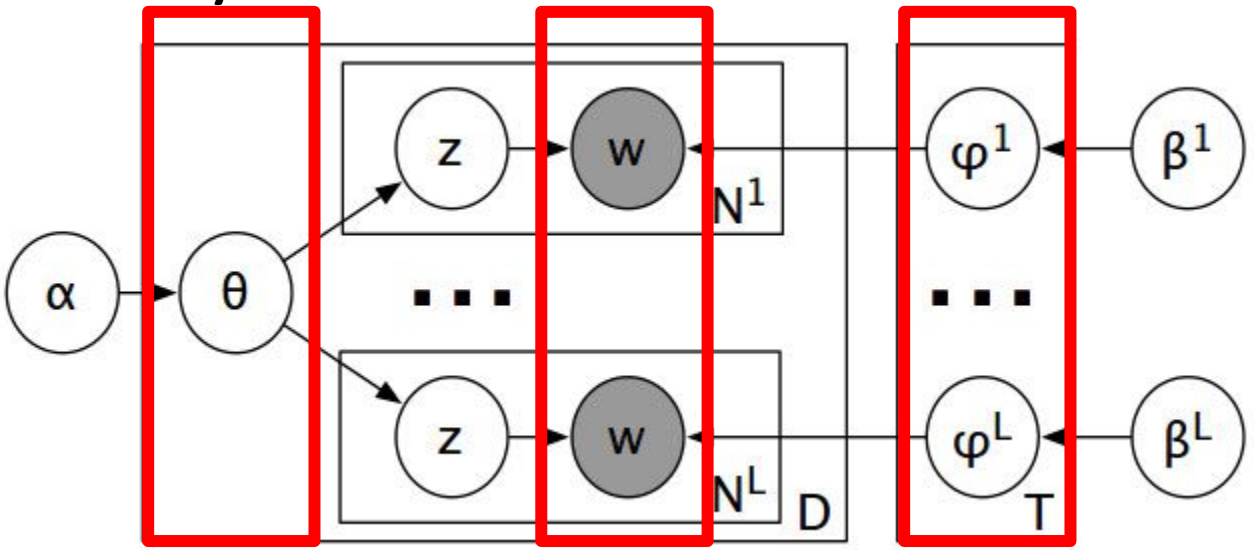
- Introduction & Dataset
- Multilingual Topic Models
- Multilingual Topics to Pseudo-parallel data
- Experimentation & Results
- Conclusion & Ongoing work

Architecture



Multilingual topic model

- PolyLDA by Mimno



Document-topic Distribution (D X **Z**)

Word-topic Distribution (V X **Z**)

Multilingual topics

TOPIC 1	
vitamin	मात्रा
quantity	विटामिन
amount	महीने
large	बड़ी
months	नाम

TOPIC 2	
clean	साफ
acid	पथरी
ulcer	अल्सर
stones	एसिड
asthma	पड़ती

TOPIC 3	
disease	रोग
blood	रक्त
heart	हृदय
diabetes	बीमारी
increases	बढ़

TOPIC 4	
cancer	कैंसर
nose	नाक
breast	शिकायत
complaint	गर्भाशय
uterus	भूख

TOPIC 1	
lake	झील
station	स्टेशन
railway	रेलवे
nearest	धीरे
munnar	मुन्नार

TOPIC 2	
famous	प्रसिद्ध
country	देश
state	प्रमुख
main	रूप
centre	राज्य

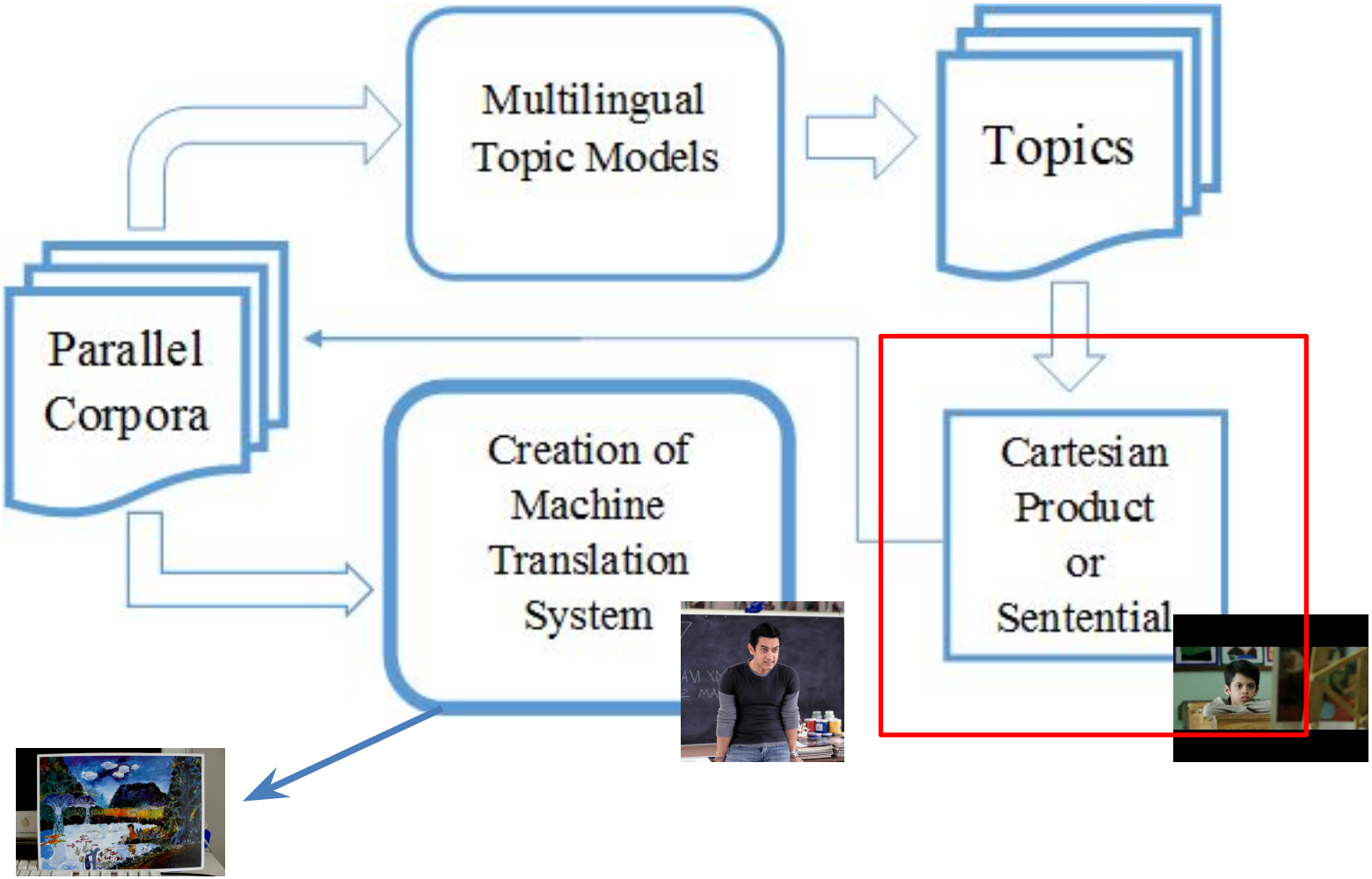
TOPIC 3	
worth	नाम
place	देखने
named	दर्शनीय
temples	नगर
city	मंदिरों

TOPIC 4	
road	मार्ग
india	भारत
route	दिल्ली
path	सड़क
kms	रोड

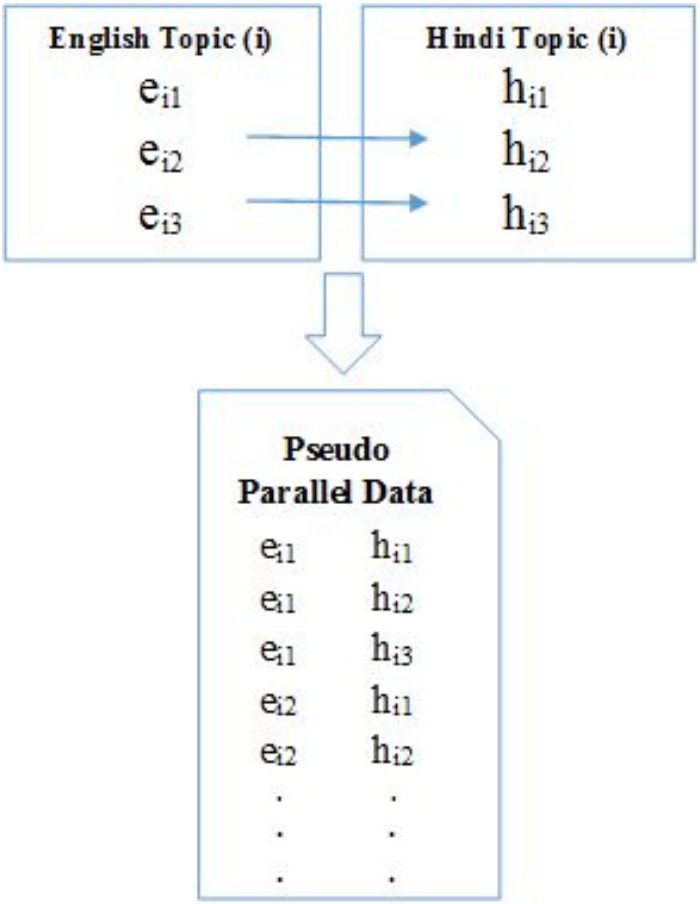
Outline

- Introduction & Dataset
- Multilingual Topic Models
- Multilingual Topics to Pseudo-parallel data
- Experimentation & Results
- Conclusion & Ongoing work

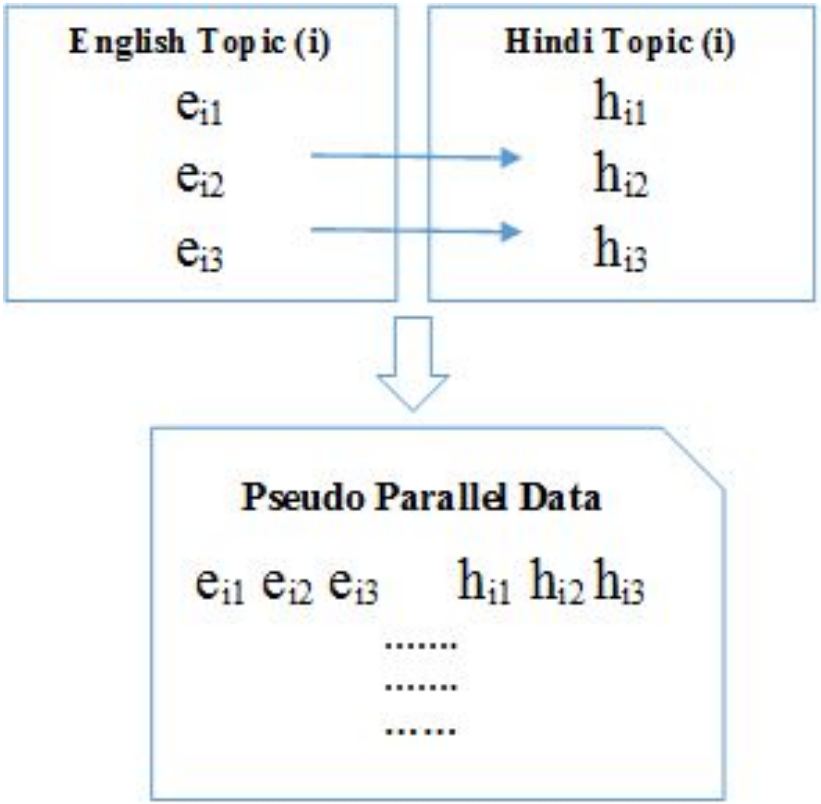
Architecture



Topics to pseudo-parallel data



Existing approach: Cartesian Product Approach



Novel approach: Sentential Approach

"Vitamin quantity amount large months"

Outline

- Introduction & Dataset
- Multilingual Topic Models
- Multilingual Topics to Pseudo-parallel data
- Experimentation & Results
- Conclusion & Ongoing work

Experimentation

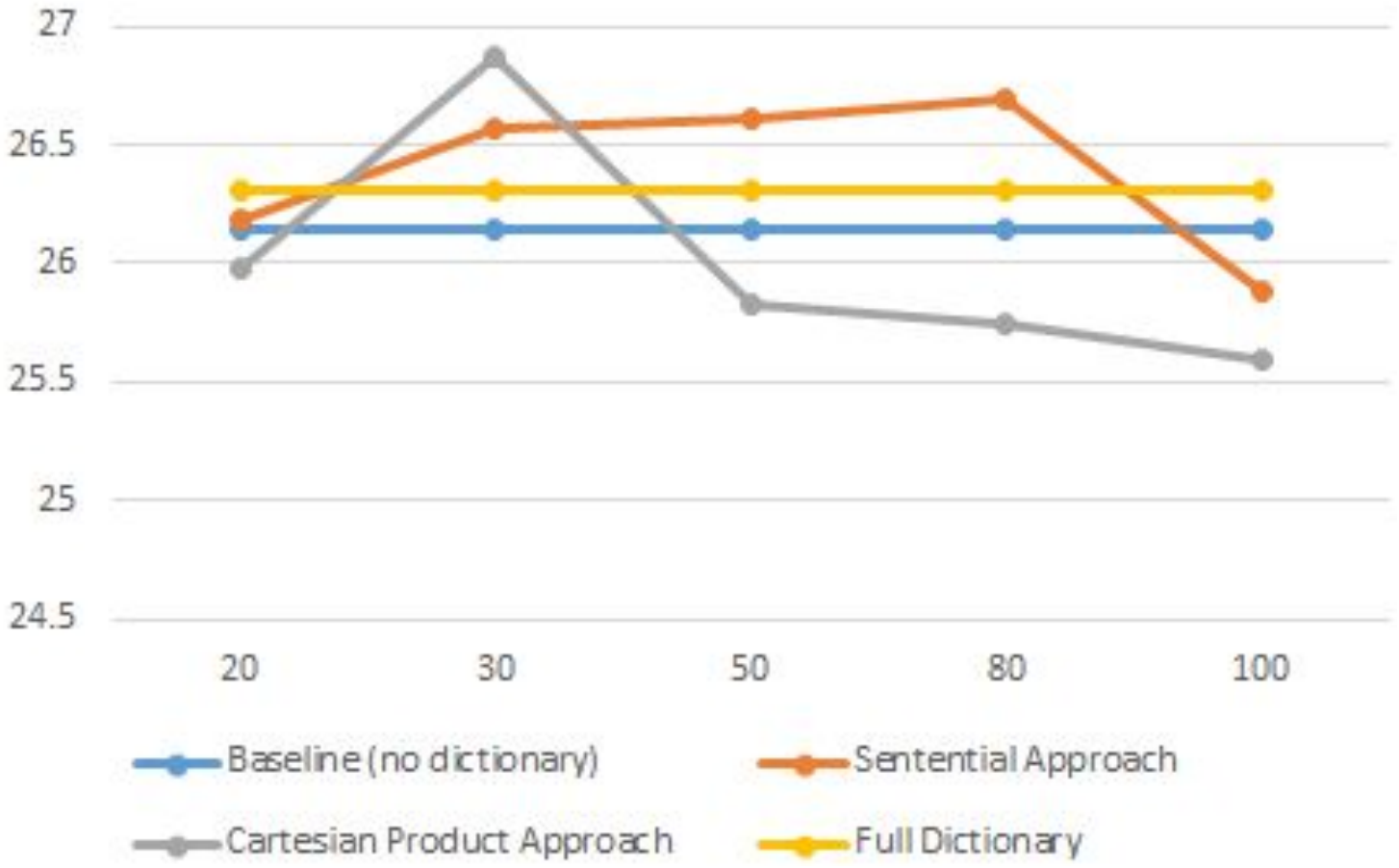
- Baseline (Only parallel corpus)
- Baseline + Pseudo-parallel data (Cartesian Product)
- Baseline + Pseudo-parallel data (Sentential Approach)
- Baseline + English-Hindi dictionary

Results (1/2): For Z = 50

	Health	Tourism
No dictionary (Base-line)	26.14	**28.68**
Cartesian product Approach (50 topics)	25.98	28.44
Sentential Approach (50 topics)	**26.25**	27.52
Full dictionary	**26.31**	**29.30**

Results (2/2): Optimizing Z

- Graph for health domain



Outline

- Introduction & Dataset
- Multilingual Topic Models
- Multilingual Topics to Pseudo-parallel data
- Experimentation & Results
- Conclusion & Ongoing work

Conclusion & Ongoing Work

- A coarse dictionary generated from multilingual topic models can be used to generate pseudo-parallel data
- With the right choice of number of topics, an improvement in BLEU score is observed

- **Ongoing work**: (a) Experimentation with multiple Indian languages, (b) Capturing inflected forms using multilingual topics

Every child, every dictionary is special.

Questions?
Comments?