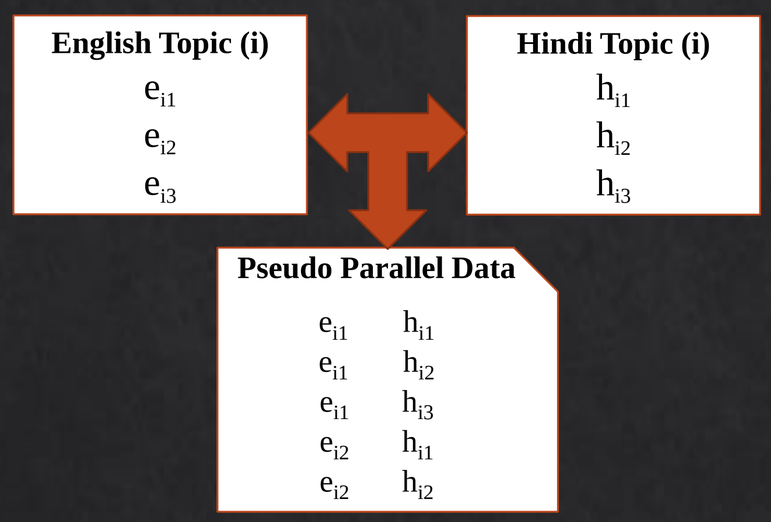
That'll do fine!: A coarse lexical resource for English-Hindi Machine Translation, using polylingual topic models

Motivation

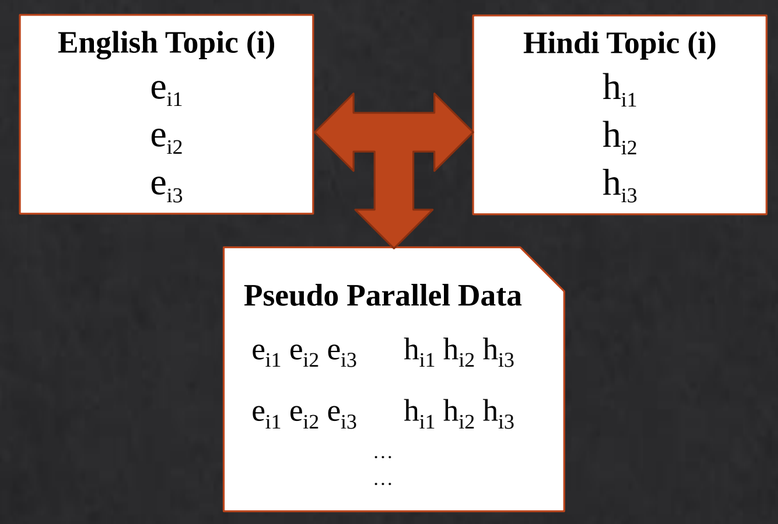
- Parallel corpora are often injected with bilingual lexical resources for improved statistical machine translation (SMT).
- Creation of such resources takes time, effort and are financially intensive.
- SMT performance is affected by the word alignment and reordering done on the training corpora.
- The previous approach (Cartesian approach) to generate pseudo-parallel data fails to provide synonymous words in parallel corpora.

Past Approach



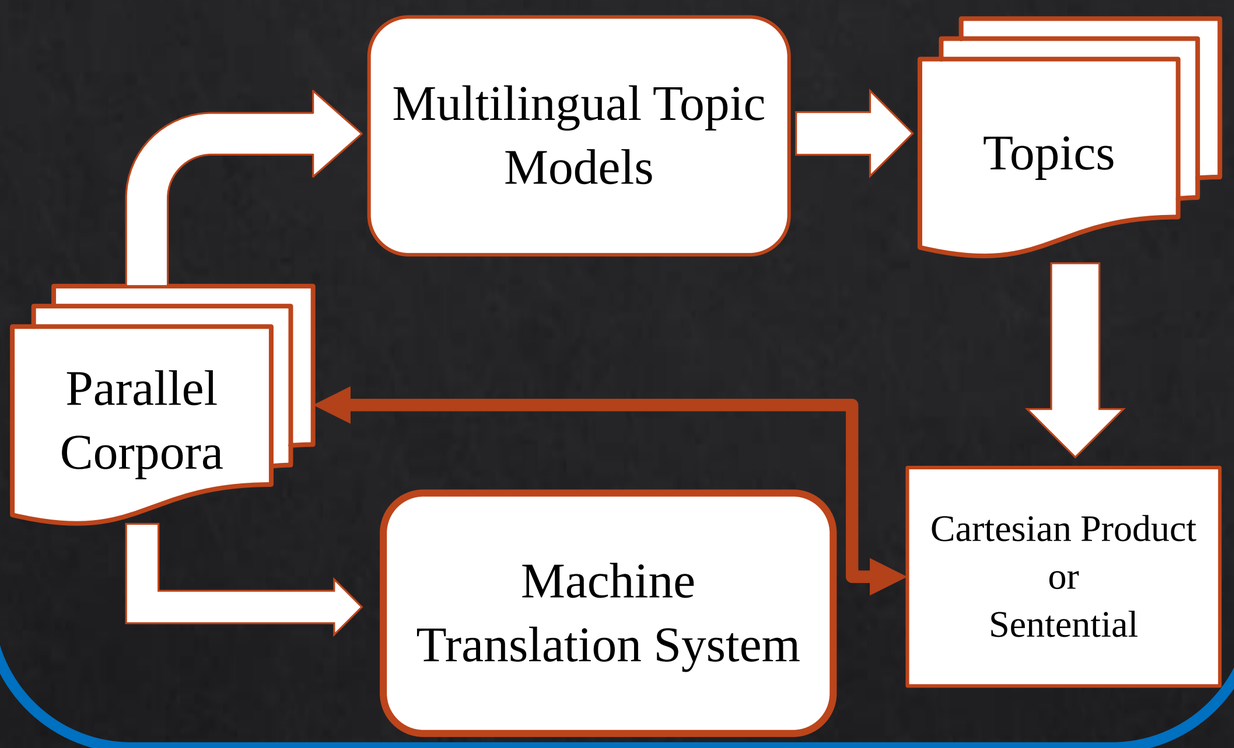
Our Approach

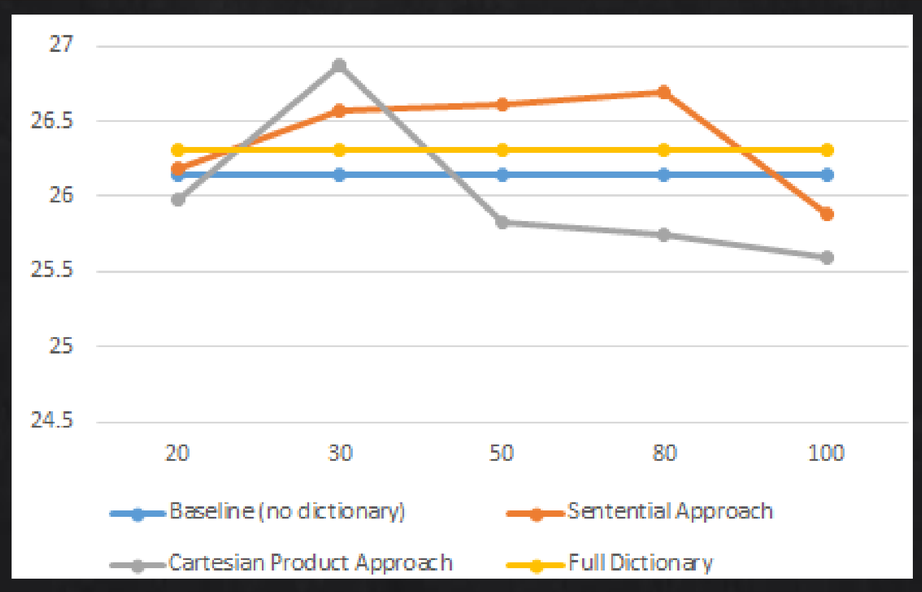
- We use unsupervised topic modelling to generate parallel topics, which can be added to the training corpora.
- We propose the sentential approach for generating pseudo-parallel data.
- Our approach aligns the pseudo-parallel data in one sentence.



- It avoids alignment of non-synonymous parallel data injection to the training corpora.
- It reduces the amount of noisy data, which was the case with Cartesian product approach.

System Architecture





System Evaluation & Results

- We evaluate our system (topic models) output quantitatively with the help of two annotators.

	Hindi	English	Kappa
A1	69.6	70.4	0.838
A2	65.6	68.4	

- We evaluate our system (topic models) output qualitatively with the help of two annotators.
  - Out of the 40 English words present in randomly chosen topics, only 7 words did not have a parallel translation.
  - Similarly, out of the 40 Hindi words, only 6 did not have translation in the corresponding topic.

(here, K=5)	Health	Tourism
No Lexical Resource	26.14	**28.68**
Cartesian Approach	25.98	28.44
Sentential Approach	**26.25**	27.52
Full lexical resource	**26.31**	**29.30**

- The results (BLEU) of four different configurations varying in terms of data which was injected in the MT system training are above.
- The graph on the left depicts a separate run of experiments where we vary the number of topics to be 20, 30, 50, 80, 100.
  - We find that our approach beats the full dictionary approach at 50 and 80 number of topics.

CFILT Centre Indian Institute of Technology for Language
भारतीय भाषा प्रौद्योगिकी केन्द्र

Diptesh Kanojia, Aditya Joshi,
Pushpak Bhattacharyya, Mark James Carman
diptesh, adityaj, pb@ cse.iitb.ac.in; mark.carman@monash.edu

MONASH University