Hindi and Marathi to English Cross Language Information Retrieval at CLEF 2007

Manoj Kumar Chinnakotla

Joint work with

Sagar Ranadive, Pushpak Bhattacharyya and Om P. Damani

Department of Computer Science and Engineering

IIT Bombay

Mumbai, INDIA

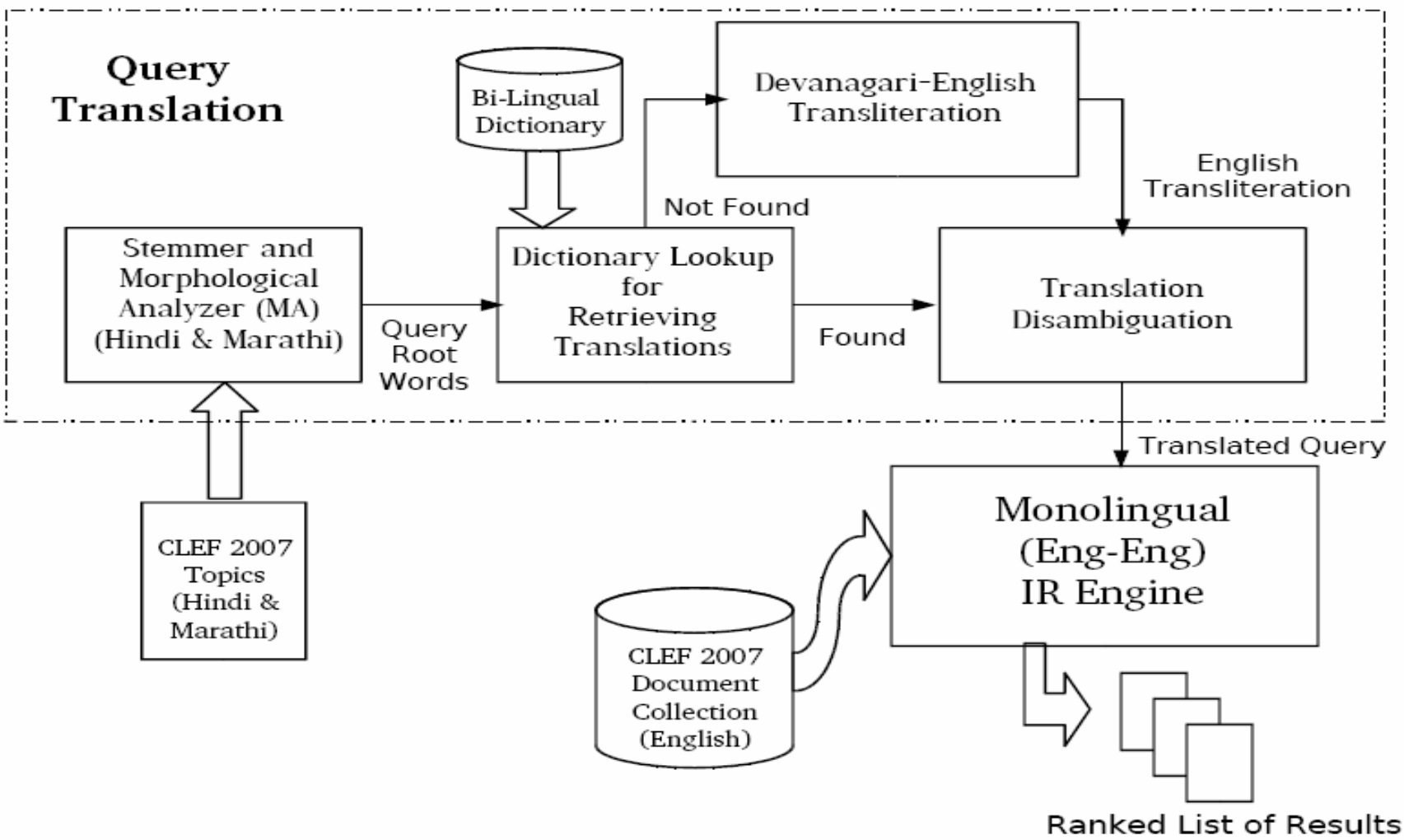
Motivation

- English still the most dominant language on the web – contributes 72% of the content
- Number of non-English users steadily rising on the web
- English penetration in India
  - Estimated to be less than 3-4%
  - Presence mostly in the urban educated sections
- CLIR systems key to enable access to English content through non-English languages

Hindi and Marathi

- Hindi
  - Official language of India
  - Spoken by almost 40% of population
- Marathi
  - Widely spoken language in Western India
  - Spoken by almost 7% of population
- Both of them
  - Written in Devanagari – A phonetic script
  - Derive vocabulary from Sanskrit

System Architecture



Language Resources

- Developed at Center for Indian Language Technologies (CFILT), IIT Bombay
- Stemmer and Morphological Analyzer
  - Rule-Based Stemmer and MA
- Bi-lingual Dictionaries
  - Hindi→English
    - 1,15,571 entries
    - Available online

    http://www.cfilt.iitb.ac.in/~hdict/webinterface_user/dict_search_user.php
  - Marathi→English
    - Relatively less coverage
    - 6110 entries

Devanagari-English Transliteration

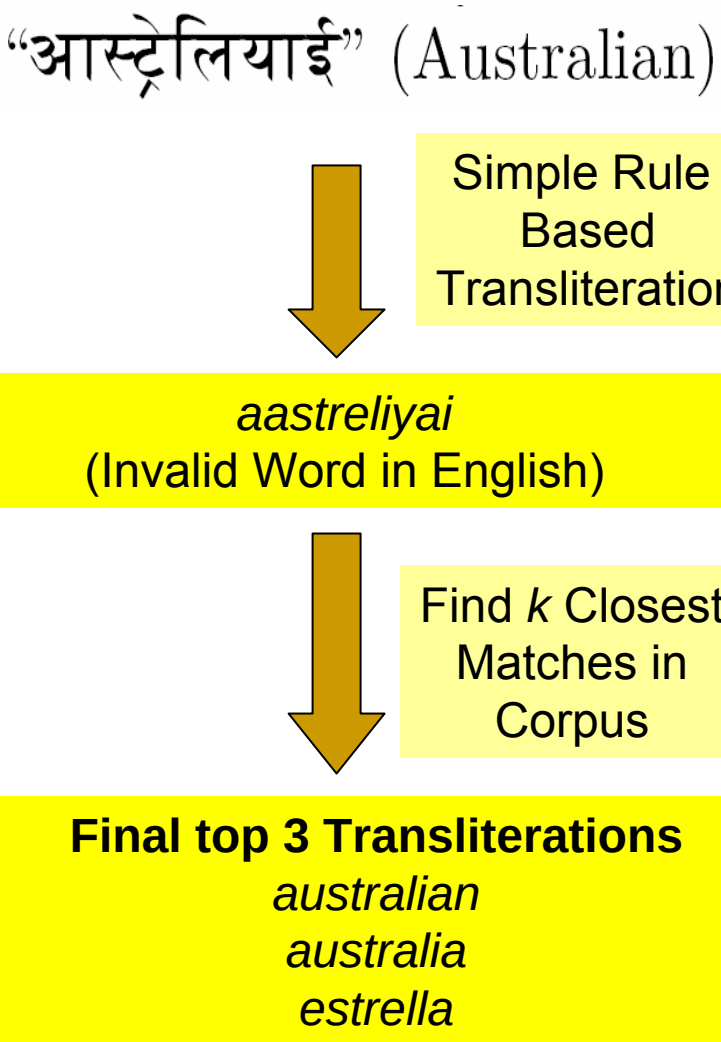
- A simple rule based transliteration scheme
- Manually created Devanagari to English transliteration mapping table for each Devanagari letter
- Given a string start from left->right and transliterate each letter using above table

Input Letter	Output String
ग़	ga
	gan
ग	ganga
ओ	gango
त्री	gangotri

Transliteration Example

Devanagari-English Transliteration (Contd..)

- Sometimes leads to invalid English words
- Resulting transliteration compared with unique words in corpus to find *'k'* closest matches
- Closeness defined in terms of string edit-distance (Levenshtein Distance)
- In current experiments, *k* set to 3



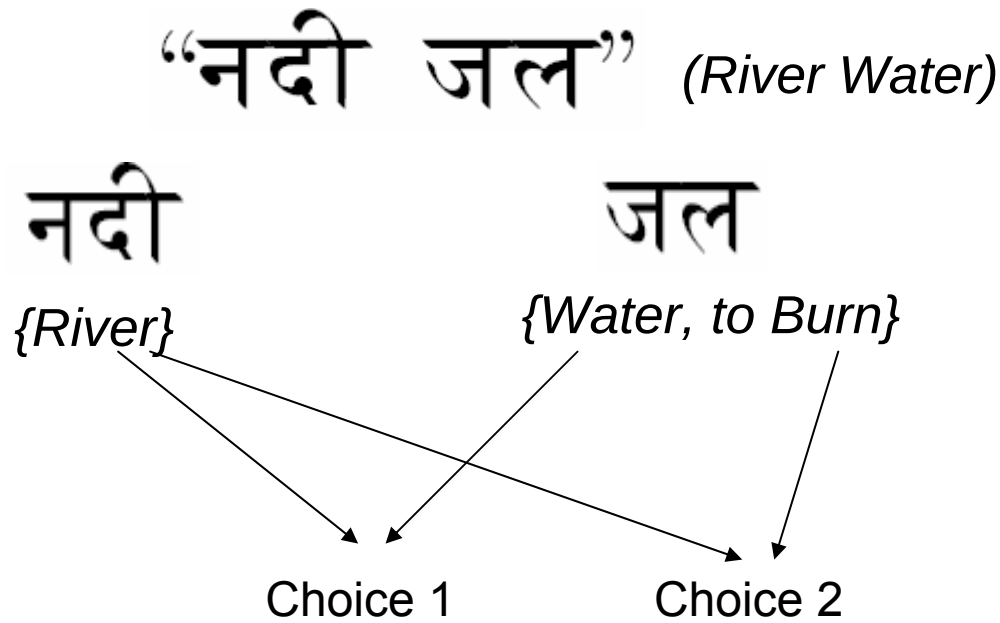
Translation Disambiguation

- Disambiguates various translation choices for each source word based word-word association measures
- For example

Hindi Query

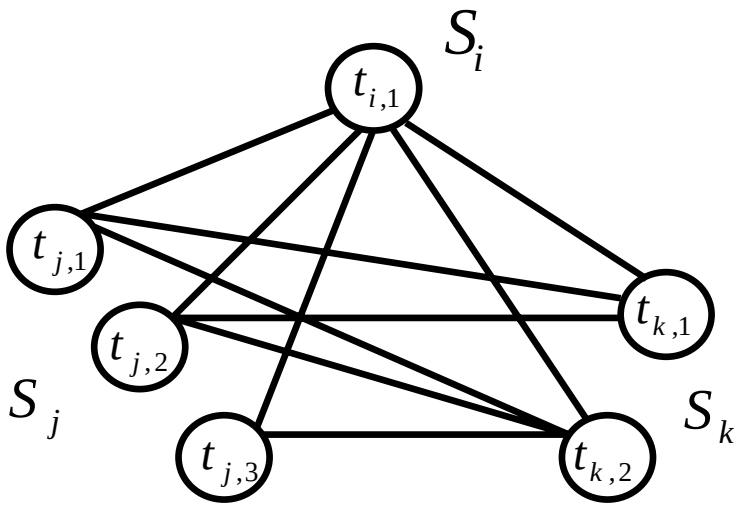
Translation Choices

Choose Based on Word-Word Association Strength



Iterative Translation Disambiguation Algorithm

- Proposed by Christof Monz *et. al.* (SIGIR 2005)
- Construct Graph
  - Nodes – Translation Choices for given source word
  - Links – Between different source word translations
- Initialize node weights assuming all translations of given source word equally likely



Iterative Translation Disambiguation Algorithm (Contd..)

- Link strength between two nodes computed based on term-term co-occurrence statistics
  - Dice Coefficient (Dice)

$$DC(t,t') = \frac{2 * freq(t,t')}{freq(t) + freq(t')}$$

  - Point-wise Mutual Information (PMI)

$$PMI(t,t') = log_2 \frac{p(t,t')}{p(t) * p(t')}$$

- The weight updation equation

$$w^n(t|s_i) = w^{n-1}(t|s_i) + \sum_{t' \in inlink(t)} l(t,t') * w^{n-1}(t'|s)$$

Previous Weight

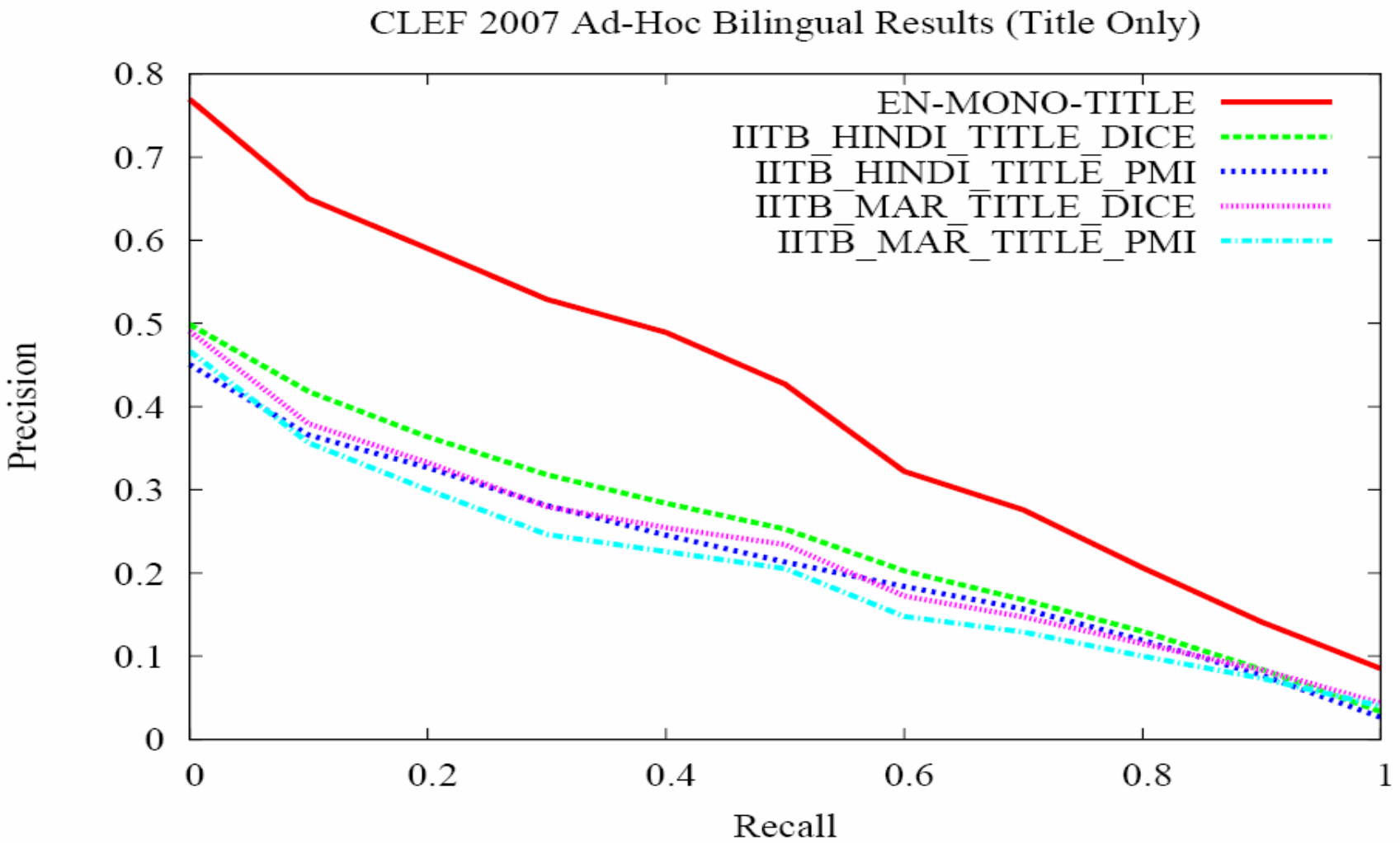
Link Strength

Weight of Neighbour

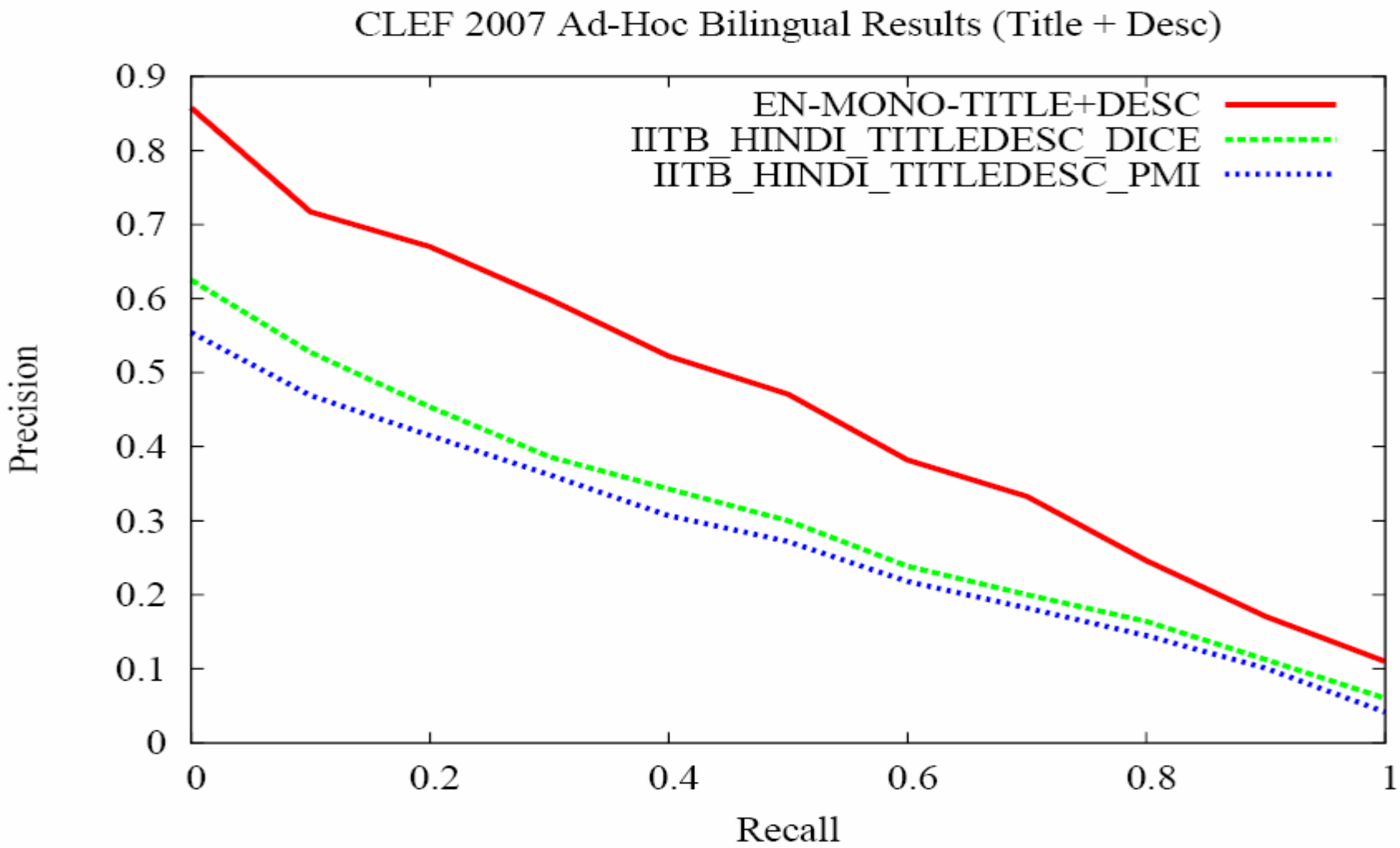
Results (Summary)

Experiment		MAP	Recall	P@20
Hindi Title	Dice	0.2366 (61.36%)	72.58% (89.16%)	0.2700 (69.05%)
	PMI	0.2089 (54.17%)	68.53% (84.19%)	0.2390 (61.12%)
Hindi Title + Desc	Dice	0.2952 (67.06%)	76.55% (87.32%)	0.3150 (73.77%)
	PMI	0.2645 (60.08%)	72.76% (82.99%)	0.2950 (69.09%)
Marathi Title	Dice	0.2163 (56.09%)	62.44% (76.70%)	0.2510 (64.19%)
	PMI	0.1935 (50.18%)	54.07% (66.42%)	0.2280 (58.31%)

Results (P-R Curves) – Title Only



Results (P-R Curves) – Title + Desc



Conclusion

- A query translation based approach taken for Hindi and Marathi to English CLIR using bi-lingual dictionaries
- Results quite encouraging – 67.06% of Monolingual baseline for Hindi, 56.09% of Monolingual baseline for Marathi
- Simple rule based transliteration taking closest edit-distance based matches from corpus performs well
- Translation disambiguation helps in selecting correct translation choices

Acknowledgements

- First author supported by the Infosys Fellowship Award
- Project linguists at CFILT, IIT Bombay
- Manish Shrivastava for help on many stemmer related issues

References

- Christof Monz and Bonnie J. Dorr, *Iterative Translation Disambiguation for Cross-Language Information Retrieval*, In SIGIR '05, Pages 520-527, New York, USA, ACM Press
- Nicola Bertoldi and Marcello Federico, *Statistical Models for Monolingual and Bilingual Information Retrieval*, Information Retrieval, 7 (1-2): 53-72, 2004
- Martin Braschler and Carol Peters, *Cross Language Evaluation Forum: Objectives, Results, Achievements,* Information Retrieval, 7 (1-2): 7-31, 2004
- Ricardo Baeza Yates and Berthier Ribeiro Neto*, Modern Information Retrieval,* Pearson Education, 2005.
- Dan Gusfield, *Algorithms on Strings, Trees and Sequences: Computer Science and Computational Biology,* Cambridge University Press, 1997.