Mapping it differently: A solution to the linking challenges

Meghna Singh, Rajita Shukla, Jaya Jha, Laxmi Kashyap, Diptesh Kanojia,
and Pushpak Bhattacharyya
Center for Indian Language Technology,
IIT Bombay

8th Global Wordnet Conference, Bucharest, 2016.

Roadmap

- Introduction
- Motivation
- Methodology
- Challenges and Solutions
- Mapping Interface
- Statistics
- Conclusions
- Future Work

Introduction

- Wordnets – online lexical resources, easily accessible, free to use, and fairly accurate.
- Applications in text processing, such as Machine Translation, Sense Disambiguation, Information Retrieval / Extraction, and NLU systems.
- Hindi Wordnet began in the year 2000 and has been used for many text processing applications for Hindi language.
- Hindi Wordnet is linked to English Wordnet via
  - Direct Linkages
  - Hypernymy Linkages

Introduction

- Wordnets – online lexical resources, easily accessible, free to use, and fairly accurate.
- Applications in text processing, such as Machine Translation, Sense Disambiguation, Information Retrieval / Extraction, and NLU systems.
- Hindi Wordnet began in the year 2000 and has been used for many text processing applications for Hindi language.
- Hindi Wordnet is linked to English Wordnet via
  - Direct Linkages
  - Hypernymy Linkages **– does not lead to accurate word / sense in all the cases.**

Motivation

- Linking synsets – parallel data between Hindi – English
- Languages are the mirrors of the societies in which they develop.
  - Specific concepts / words, geographical and cultural milieus seep in.

Motivation

- Linking synsets – parallel data between Hindi – English
- Languages are the mirrors of the societies in which they develop.
  - Specific concepts / words, geographical and cultural milieus seep in.
  - Such ‘concepts’ have no parallel words to map!

Motivation

- Linking synsets – parallel data between Hindi – English
- Languages are the mirrors of the societies in which they develop.
  - Specific concepts / words, geographical and cultural milieus seep in.
  - Such ‘concepts’ have no parallel words to map!
- When two languages which are set apart as Hindi and English, have to be linked at the conceptual level, it is bound to throw up the challenge of lexical and conceptual gaps.

Motivation

- Linking synsets – parallel data between Hindi – English
- Languages are the mirrors of the societies in which they develop.
  - Specific concepts / words, geographical and cultural milieus seep in.
  - Such ‘concepts’ have no parallel words to map!
- When two languages which are set apart as Hindi and English, have to be linked at the conceptual level, it is bound to throw up the challenge of lexical and conceptual gaps.
- To overcome such challenges, the two kinds of linkages were being created.

Motivation

- Direct Linkage
  - गंधयुक्त (*gandhayukta*) which means,
    - जो गंध से युक्त हो
    - which fragrance with is
    - *jo gandha se yukta ho*
    - which has fragrance

Motivation

- Direct Linkage
  - गंधयुक्त (*gandhayukta*) which means,
    - जो गंध से युक्त हो
    - which fragrance with is
    - *jo gandha se yukta ho*
    - which has fragrance,
  - can be directly linked to,
  - '**odorous**', which means having a natural fragrance.

Motivation

- Hypernymy Linkage

Motivation

- Hypernymy Linkage – adopted from EuroWordnet.

Motivation

- Hypernymy Linkage – adopted from EuroWordnet.
- ‘Something is better than nothing’ – a super-ordinate synset leading to a vauge description of the concept.
-

Motivation

- Hypernymy Linkage – adopted from EuroWordnet.
- ‘Something is better than nothing’ – a super-ordinate synset leading to a vauge description of the concept.
- सदावर्त (*sadaavarta*)
  -

Motivation

- Hypernymy Linkage – adopted from EuroWordnet.
- 'Something is better than nothing' – a super-ordinate synset leading to a vauge description of the concept.
- सदावर्त (*sadaavarta*)
  - the act of distributing food and other essential items to poor people for a specific time period according to a vow undertaken.

Motivation

- Hypernymy Linkage – adopted from EuroWordnet.
- 'Something is better than nothing' – a super-ordinate synset leading to a vauge description of the concept.
- सदावर्त (*sadaavarta*)
  - the act of distributing food and other essential items to poor people for a specific time period according to a vow undertaken.
  - But, can be linked to
  - *Charity - 'an activity or gift that benefits the public at large'*

Motivation

- However, Concepts like
-
-

Motivation

- However, Concepts like
- छेदन (*chhedan*) , which means

  छेद करने की क्रिया
  - make hole act
  - *chhed karane kii kriyaa*
  - act of piercing.

Motivation

- However, Concepts like
- छेदन (*chhedan*) , which means

  छेद करने की क्रिया
  - make hole act
  - *chhed karane kii kriyaa*
  - act of piercing.
- In the absence of a matching synset in the same POS category in the English WordNet,

Motivation

- However, Concepts like
- छेदन (*chhedan*) , which means

  छेद करने की क्रिया
  - make hole act
  - *chhed karane kii kriyaa*
  - act of piercing.
- In the absence of a matching synset in the same POS category in the English WordNet,
- It was linked to "*deed*" - *something that people do or cause to happen*.
-

Motivation

- However, Concepts like
- छेदन (*chhedan*) , which means

  छेद करने की क्रिया
  - make hole act
  - *chhed karane kii kriyaa*
  - act of piercing.
- In the absence of a matching synset in the same POS category in the English WordNet,
- It was linked to "*deed*" - *something that people do or cause to happen*.
- Very vauge, far fetched linkages.
- WSD ->MT (gap increases)

Methodology

- Translation or Transliteration of the synset members, and,

Methodology

- Translation or Transliteration of the synset members, and,
- Translation of the gloss of the concept

Methodology

- Translation or Transliteration of the synset members, and,
- Translation of the gloss of the concept.
- We search various lexical resources, look for valid usages on the internet / in a corpus (frequency counts, corpus occurrences etc.)

Methodology

- Translation or Transliteration of the synset members, and,
- Translation of the gloss of the concept.
- We search various lexical resources, look for valid usages on the internet / in a corpus (frequency counts, corpus occurrences etc.)
- As far as possible, we try to avoid coining words / terms for the purpose.

Methodology

- Translation or Transliteration of the synset members, and,
- Translation of the gloss of the concept.
- We search various lexical resources, look for valid usages on the internet / in a corpus (frequency counts, corpus occurrences etc.)
- As far as possible, we try to avoid coining words / terms for the purpose.
  - But then, exceptions have to made.
- The mappings are made via an online interface (described later)

Methodology

- Translation or Transliteration of the synset members, and,
- Translation of the gloss of the concept.
- We search various lexical resources, look for valid usages on the internet / in a corpus (frequency counts, corpus occurrences etc.)
- As far as possible, we try to avoid coining words / terms for the purpose.
  - But then, exceptions have to made.
- The mappings are made via an online interface (described later)
- Users can see the mappings on the online interface of Hindi Wordnet.
  - In case a Hypernymy linkage exists for a current synset, and is too far fetched, we remove it, and instead use Bilingual Mappings

Challenges and Solutions

- Creation of mappings had to be divided into four major categories, based on the problems faced.

Challenges and Solutions

- Creation of mappings had to be divided into four major categories, based on the problems faced.
  - **Words / Concept not available in English WordNet.**
  -

Challenges and Solutions

- Creation of mappings had to be divided into four major categories, based on the problems faced.
  - **Words / Concept not available in English WordNet.**
  - **Required sense missing in the English WordNet.**
  -

Challenges and Solutions

- Creation of mappings had to be divided into four major categories, based on the problems faced.
  - **Words / Concept not available in English WordNet.**
  - **Required sense missing in the English WordNet.**
  - **Culture Specific Words.**

Challenges and Solutions

- Creation of mappings had to be divided into four major categories, based on the problems faced.
  - **Words / Concept not available in English WordNet.**
  - **Required sense missing in the English WordNet.**
  - **Culture Specific Words.**
  - **Language Specific Words.**

Challenges and Solutions

- **Words / Concept not available in English WordNet.**
- Transliteration,
  - When no suitable word in the English WordNet is found to represent the Hindi concept, we transliterate the word and translate the gloss accordingly. For example, पदयात्रा (*padayatra*)
  - which means,
  - *- Kisii vishesh uddeshya (visheshkar raajnaitik yaa dhaarmik) se paidal kii jaane walii yaatraa*
    - a foot journey undertaken for some special purpose (especially political or religious).
  - Was initially hypernymy linked to "Hike" – a long walk usually for exercise or pleasure.

Challenges and Solutions

- **Words / Concept not available in English WordNet.**
- Translation,
- the synset members are translated along with the gloss in English.
- An example is अप्सरा (apsaraa)
  - which means,
  - - *Swarga meM Indra kii sabhaa meM naachane-gaane walii sundariyaan*
    - beautiful ladies who dance and sing in Indra's court in the heaven.
  - Was earlier linked to 'nymph' - *a minor nature goddess usually depicted as a beautiful maiden*.
  - Now. Translated to "Celestial Dancer" - *beautiful ladies who dance and sing in heaven in the court of Indra.*

Challenges and Solutions

- **Required sense missing in the English WordNet**
- the synset is present in the English WordNet, but the given sense/s does not match the one required for the Hindi synset
- For example, **फूँकना** (*phuunkana*)
- which means,
- *- phuunk maar kar dahakaanaa yaa prajjawalit karanaa*
  - to light or inflame by blowing.
- Although, the English WordNet has four senses of the word *ignite* but this particular sense is not present.
- So, we assign the word "*ignite" and* translated the gloss as *cause to start burning by exhaling hard through mouth.*

Challenges and Solutions

- **Culture Specific Words**
- words specific to Indian culture and hence not found in the English WordNet

Challenges and Solutions

- **Culture Specific Words**
- words specific to Indian culture and hence not found in the English WordNet
- For example, बिछिया (*bichhiyaa*), which means
  - *pair kii ungaliyoM meM pahanane kaa chhallaa*
  - ring worn on toes.
- It has a hypernymy linkage to *jewelry*, which means *an adornment (as a bracelet or ring or necklace) made of precious metals and set with gems (or imitation gems*).
- But, It does not convey the meaning accurately.
- "Toe Ring" is very commonly used word to represent the concept, hence, was included as a mapping.

Challenges and Solutions

- **Language Specific Words**
- There are many words in Hindi Wordnet which capture the peculiar grammar of the language. It is but natural that their counterparts will not be available in English. Hence, these words require bilingual mappings.
-

Challenges and Solutions

- **Language Specific Words**
- There are many words in Hindi Wordnet which capture the peculiar grammar of the language. It is but natural that their counterparts will not be available in English. Hence, these words require bilingual mappings.
  - **Idiomatic Expressions**
    - They are highly culture specific and so they require special treatment, becoming perfect candidates for bilingual mapping, specifically those not available in English.
  - **Causative Verbs**
    - Causative verbs indicate an action that the subject does not directly perform, but rather causes to happen, perhaps by causing some other agent to perform the action. Such verbs are a well-known feature of Hindi and are represented in English as a phrase.
  - **'Be' Form of Conjunct Verbs**
    - Many conjunct verbs have corresponding intransitive forms which employ होना (to be).

Challenges and Solutions

- **Language Specific Words**
- There are many words in Hindi Wordnet which capture the peculiar grammar of the language. It is but natural that their counterparts will not be available in English. Hence, these words require bilingual mappings.
  - **Idiomatic Expressions**
    - They are highly culture specific and so they require special treatment, becoming perfect candidates for bilingual mapping, specifically those not available in English.
  - For Example, हाथ खुला होना (*haath khula honaa*), which would literally mean "to have an open hand", but the idiomatic sense is
    - *daan, vyaya, aadi ke saMbandha meM udaar pravritti honaa.*
    - to be of generous tendency towards donation, expenditure, etc.

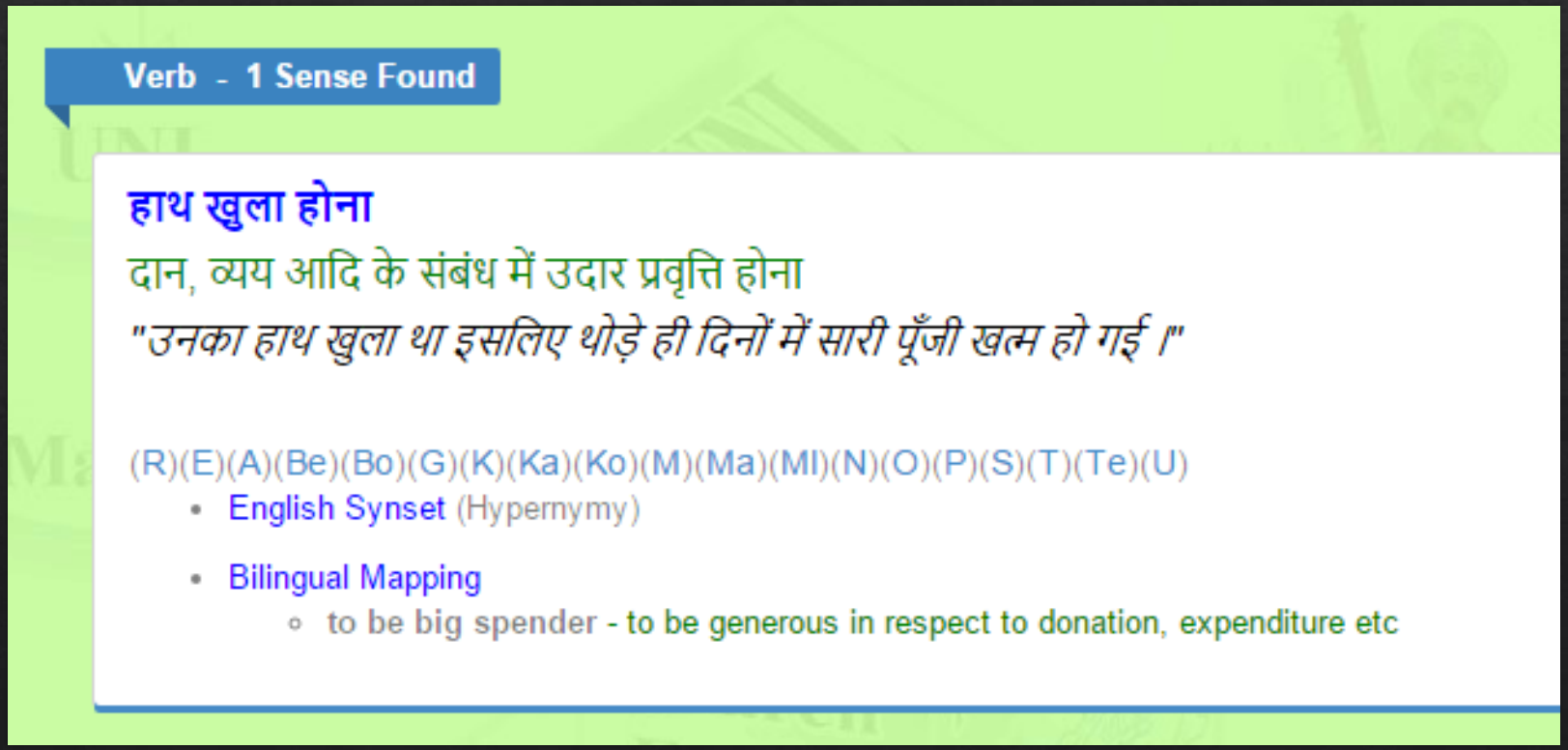
Challenges and Solutions

- **Words for which Hypernymy Relations are unavailable**
- Wordnet does not have hypernymy relation for adjectives and adverbs. Thus these words in the Hindi wordnet when not linked to direct English words, do not have an option of hypernymy linkage. In such cases, they have to be invariably given bilingual mappings.
-

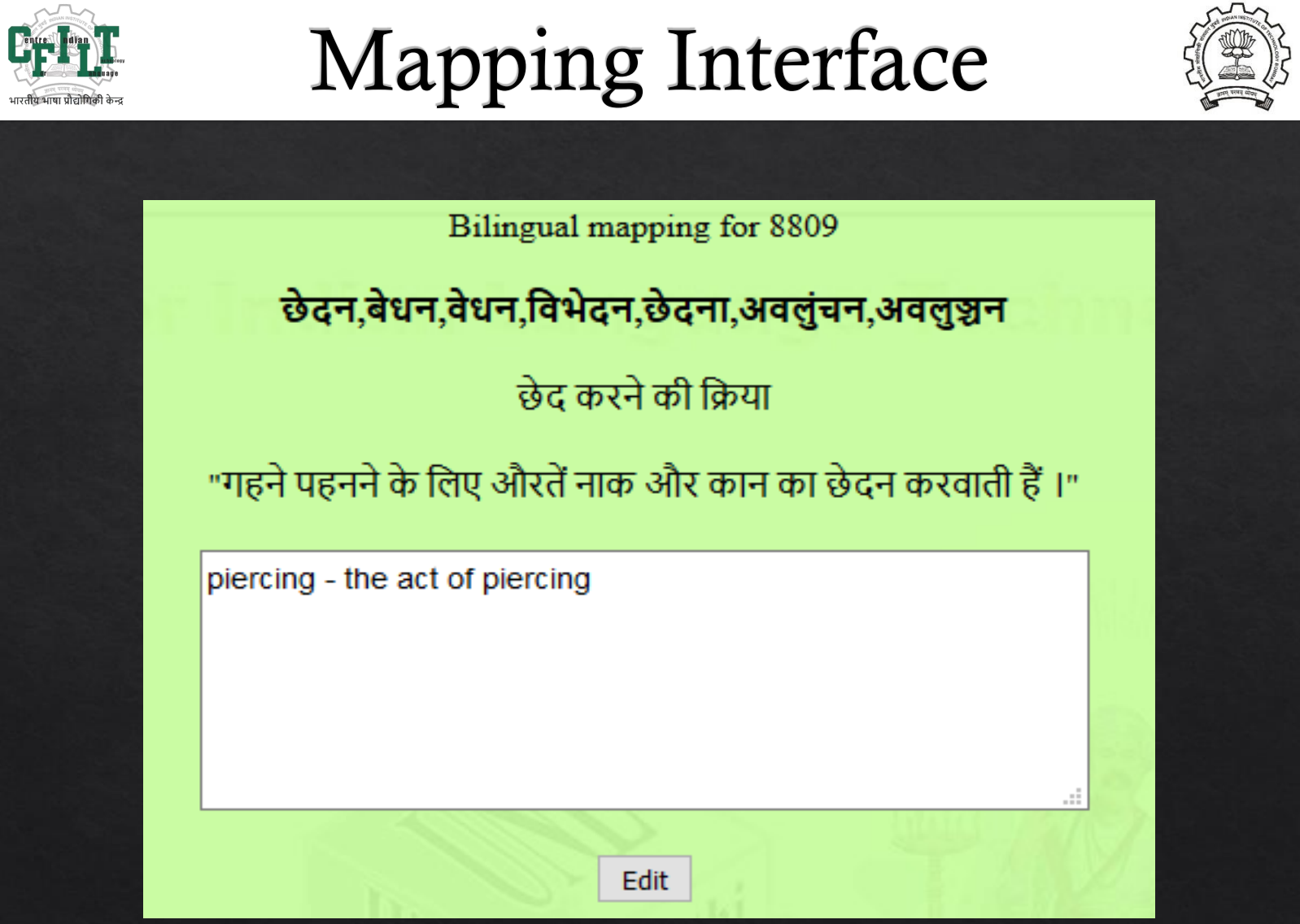
Challenges and Solutions

- **And also, "Proper Nouns"**
- Hindi Wordnet has more than 16,000 proper nouns, most of which are names of persons, places and organizations specific to India. All such words could not have been given a place in the English WordNet, making linkage difficult. Initially they were given hypernymy linkages to very distant synsets.
  - Characters from Indian Epics, or Mythology
  - Indian Leaders etc.

Mapping Interface



Mapping Interface



Statistics

Part of Speech	Total Synsets in HWN	Direct Linkage	Hypernymy Linkage	Biligual Mappings	Total Linkage
Noun	29070	11582	8184	2110	21876
Adjective	6171	3541	0	331	3872
Verb	3303	1992	207	129	2328
Adverb	475	343	0	27	370
Total	39019	17458	8391	**2597***	28446

*Continously being updated!

Conclusions

- Hindi – English linkages are an effort in the direction of improving cross lingual WSD, MT, and other NLP applications.
- Direct or Hypernymy linkages, clearly, do not suffice the need of mapping most / all of the concepts in an accurate manner.
- We introduce bilingual mappings for Hindi – English language pair, to map the concepts accurately.
- An online linking facility for mapping them has been created.
- More than 4000 Synsets (current statistics) have been mapped.
- An evaluation as to how these are helpful, shall be performed soon.

Acknowledgements

- Department of Electronics and Information Technology, Ministry of Communications and IT, Government of India.
- IRCC, IIT Bombay for the Travel Fund.

Thank you!

- Thank you for all you attention.
- Questions ?
- Suggestions ?