Challenges in Multilingual Domain-Specific Sense-marking

Jaya Saraswati, Rajita Shukla, Sonal Pathade, Tina Solanki,
Pushpak Bhattacharyya
Department of Computer Science and Engineering
Indian Institute of Technology, Bombay
Powai, Mumbai – 400076
Maharashtra, India
{jayas, rajita, sonal, thsolanki, pb}@cse.iitb.ac.in

Abstract

Annotation plays a key role in today's NLP scenario and this paper discusses challenges involved in one of the toughest annotation tasks - sense marking. In an effort to train the machine to understand the written language and thus to ensure speedy and high-quality translation, a huge amount of data needs to be sense-marked accurately by humans using an authentic and standard lexicon. In the work reported here, the corpus is taken from tourism domain and the Princeton wordnet (Version 2.1) is used as the sense inventory for English text while the Hindi and Marathi wordnets have been used for Hindi and Marathi texts respectively. A word may have a number of senses and in identifying which particular sense has been used in the given context, word sense disambiguation becomes a critical necessity. The corpus was independently tagged by different sense-markers and it was found that the inter annotator agreement on word sense disambiguation was about 80 % across the three languages, *i.e.*, English, Hindi and Marathi. Though the sense distinctions in the wordnets are quite fine-grained, there have been cases when the senses provided there have been inadequate and the human sense-markers have faced problems. The study records such challenges and their handling.

Keywords: sense-marking, wordnet, tourism, word sense disambiguation, culture-specific, challenge, inter annotator agreement.

1 Introduction

The famous Princeton University English wordnet[1], an electronic lexical database, has paved the way for other wordnets in different languages across the world. It set the design for having the nouns, verbs, adjectives and adverbs of a language grouped under sets of synonyms, or synsets[2]. Apart from functioning as a dictionary and thesaurus combined into one, it is used greatly in automatic text analysis and artificial intelligence applications.

Hindi[3] and Marathi[4] wordnets have been developed by researchers at the Center for Indian Language Technology, Computer Science and Engineering Department, IIT Bombay. Similar in design to the Princeton wordnet for English, Hindi wordnet incorporates additional features to capture the complexities of Hindi. Since Marathi wordnet is based on the Hindi wordnet, it directly inherits the IDs and semantic relations of words from there.

While working towards building a Machine Translation system from English to any Indian language, word sense ambiguity has been a prominent issue[5]. In a given text, the occurrence of a particular word will correspond to only one sense and nearby words provide strong and consistent clues to the sense of a target word.

The roadmap of the paper is as follows: Section 2 describes the methodology for sense-marking, the sense-marker tool, a description of how it works, and also the screenshot of this tool. Section 3 and all its subsections describe the options for sense-marking that have been considered along with examples to illustrate the point.

[1] http://wordnetweb.princeton.edu/perl/webwn/

[2] Fellbaum, C. 1998. *Wordnet: An Electronic Lexical Data base.* The MIT Press

[3] http://www.cfilt.iitb.ac.in/wordnet/webhwn/

[4] http://www.cfilt.iitb.ac.in/wordnet/webmwn/

[5] Mitesh Khapra, Sapan Shah, Piyush Kedia and Pushpak Bhattacharyya, *Projecting Parameters for Multilingual Word Sense Disambiguation*, Empirical Methods in Natu ral Language Processing (EMNLP09), Singapore, August, 2009

Section 4 presents some other challenges faced by the sense-markers. Section 5 presents such cases where inadequacies of the English, Hindi and Marathi wordnets were encountered. The 6th section shows a comparative study of sense marking and the other annotation tasks. The 7th and the final section winds up the discussion by presenting the conclusions and future work on the issues.

2 Methodology followed in Sense Marking

In the process of training the machine to disambiguate a word sense for proper translation in the Tourism domain we use a sense-marking tool[6]. It is a software tool developed to provide the lexicographers with an easy and efficient way of sense tagging the words. A Graphical User Interface based tool using Java facilitates the task of manual sense marking.

The SenseMarker Interface is shown in figure 1 below.

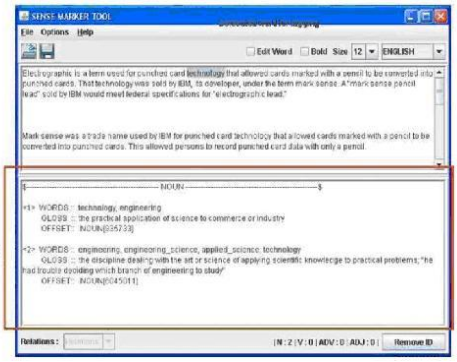

Figure 1: Screenshot of the SenseMarker tool

The tool supports nine languages including English[7], Hindi[8], Marathi[9], Tamil[10], Telugu[11], Kannada[12], Malayalam[13], Bengali[14] and Punjabi[15]. It displays the different senses of a word (as available in the wordnet) along with some other useful information like the gloss and entries of each Synset to which the word belongs. It allows the user to select the correct sense of the word from amongst all the senses. The word can be tagged by just a single click on the correct Synset.

The Steps to Sense-mark a document:

1. The sense-marker sets the Language for which Sense Tagging is to be done.
2. Open the file by clicking on the Open MenuItem of the File Menu or the Tool Bar Open Icon or by pressing CTRL+ O on the key board.
3. A File Chooser menu gets opened. The user has to select the file for tagging and press the open button of the same or double click on the file.

---

[6] Prof. Pushpak Bhattacharyya, Shashank Chauhan, Soumya Nair: Sense Marker Tool : Guide To Sense Marker Tool (B.Tech project report, 2008)

[7] English is the second official language in India. At the time the constitution entered into force, English was used for most official purposes both at the federal level and in the various states The constitution envisaged the gradual phasing in of local languages, principally Hindi, to replace English over a fifteen-year period, but gave Parliament the power to, by law, provide for the continued use of English even thereafter. Accordingly, English continues to be used today, in combination with Hindi (at the central level and in some states) and other languages (at the state level).

[8] Hindi/Khadi boli belongs to the Indo-Aryan language subgroup of Indo-European language family.It is a dialect continuum of the Indic language family in the northern plains of India. 2001 census of India noted 422,048,642 speakers of this language. It is spoken in the Indian states and union territories of Bihar,Chhattisgarh, Delhi, Haryana, Himachal Pradesh, Jharkhand, Madhya Pradesh, Rajasthan, Uttar Pradesh and Uttarakhand.

[9] Marathi is an Indo-Aryan language spoken by the Marathi people of south western India and is the official language of the state of Maharashtra. 2001 census of India noted 71,936,894 speakers of this language.

[10] Tamil is the only surviving Classical language in the world and is a Dravidian language. According to the 2001 census of India there are 60,793,814 speakers of this particular language.

[11] Telugu is a Dravidian language mostly spoken in the Indian state of Andhra Pradesh. According to the 2001 census of India there are 74,002,856 speakers of this particular language.

[12] Kannada is one of the major Dravidian languages of India, spoken predominantly in the state of Karnataka. 2001 census of India recorded 37,924,011 speakers of this language.

[13] Malayalam is one of the four major Dravidian languages of South India. According to the 2001 census of India there are 33,066,392speakers of this particular language.

[14] Bengali or Bangla is an Indo-Aryan language of the eastern Indian subcontinent, evolved from the Magadhi Prakrit and Sanskrit languages. Bengali is native to the region of eastern South Asia known as Bengal, which comprises present day Bangladesh, the Indian state of West Bengal, southern Assam - also known as Barak Valley, and part of Tripura. With nearly 230 million total speakers, Bengali is one of the most spoken languages (ranking fifth or sixth) in the world.

[15] Punjabi is spoken as native language by 44.15% of Pakistanis and 2.85% of Indians is an Indo-Aryan language spoken by inhabitants of the historical Punjab region (in Pakistan and India).According to the Census of India 2001; there are 29,102,477 Punjabi speakers in India.

4. The file/document gets opened in the Tagging Window with added new lines at the end of each statement to demarcate the sentences properly and readability is increased.
5. The word, to be sense tagged, should be single clicked and for a compound word the user is required to drag to select.
6. The synsets for the selected word are displayed in the Synset Window.
7. To assign sense (POS + Offset) to the word the user has to click on the respective synset which makes the right sense and the offset and the pos gets assigned to the word.

However, there have been occasions when this task of sense marking was not a simple one. It was found that either the sense was not present in the wordnets or the existing sense was not sufficient or the components of the compound expressions could not be separately disambiguated to provide the correct sense. We now proceed to discuss these issues.

3 Options for Sense-marking

In light of the things mentioned above, the sense markers had the following options:

a. Marking the word with the exact sense
b. Marking with a subsuming sense
c. Marking with the closest sense
d. Marking with the exact sense even if the sense does not mention the particular word as a synset member
e. Creating a new sense

3.1 Marking the word with the exact sense

This is the ideal and most desirable situation. It is the task of the sense-marker to assign senses to as many words as possible. When the word is available in the sense repository with its complete and correct set of senses, the sense marker essentially has to apply her/his knowledge of language and the sense of context to assign the sense accurately.

3.2 Marking with a subsuming sense (using hypernymy)

When the exact sense was not found, sometimes a subsuming sense was tagged as it contained the essence of the original word. For example, the equivalent of various words in Hindi which denote a *container* or a *cooking pot* are not there in English, but they could be tagged to vessel or pan respectively.

3.3 Using a close sense

The basic observation was that in the absence of an exact match of the sense for a given word, it was decided that a close or nearby sense should be tagged. The word *festival*, for example, shows two senses in the wordnet, (a) a day or period of time set aside for feasting and celebration, and, (b) an organized series of acts and performances (usually in one place); "a drama festival". In the Indian context, a sense like *An occasion for feasting or celebration, especially a day or time of religious significance that recurs at regular intervals*[16] looked more appropriate, but since the wordnet sense (a) did not look too out of place, it was tagged.

3.4 Marking with the exact sense even though the sense does not mention the particular word as a synset member

If the exact sense was found in some other existing synset of the wordnet, then the word was made a part of that synset. It was decided that such words should be enclosed with a hash mark followed by the ID of that synset. Quite a few words fell in this category. For example:

a. *Ganga*: This word was tagged to the concept (an Asian river; rises in the Himalayas and flows east into the Bay of Bengal; a sacred river of the Hindus) where the existing synset was Ganges, Ganges River. The ID of this synset was given to the word Ganga as, by right, it should have been a member of this synset. The tagging looked like this - **#Ganga#_9153625.**

b. *Pulao*: The concept was same as that of pilaf, pilaff, pilau, pilaw (rice cooked in well-seasoned broth with onions or celery and usually poultry or game or shellfish and sometimes tomatoes), and so it was tagged with this ID.

Another instance was when the concept was present in the wordnet but there was a different word for it in Indian English. For example, the word *raw* as in *raw mangoes* where it conveyed the sense of being *green, unripe, unripened, immature*, (the gloss being – *not fully developed or mature; not ripe*) could be directly traced to this concept. A similar case was with the concept of

[16] www.thefreedictionary.com

salad leaves, the way the word *lettuce* is used in Indian English.

3.5 Creating a new sense

This option was adopted on occasions where a word appears in the document, the sense of which is either present in the wordnet but is not appropriate in the context or is completely absent from the wordnet. This is clearly obvious in cases of culture-specific word entities. It was decided that a new sense should be created for them and stored in the local copy of the wordnet. All such words were enclosed between an opening # and a closing # symbol. A script to parse the words between these symbols would be written and new synsets for these words would be created. The synset IDs will start from 200000. The categories for these are:

Culture-specific words

Specific words from all cultures cannot be stored in any lexicon. There are certain terms which express a concept which is not universal in nature. They are embedded in the culture of that particular land where they originated. Such words are not sufficiently present in the wordnets. These concepts pertain more specifically to places which were European Colonies at some point of time. However, since English is the most important language of communication all over the world, the sense inventories are expected to contain most of them. In the Indian context, the Princeton wordnet at times does come up to this expectation and has captured little-known concepts as well. An interesting instance is that of the word *Raita* which has the gloss *an Indian side dish of yogurt and chopped cucumbers and spices* in the English wordnet. This is a common North Indian dish which is not very familiar elsewhere in India itself and yet it has found its way in an English lexicon. Other English dictionaries have also incorporated many Indian words like *guru, Brahmin, chapatti, gherao, etc*. Yet the fact remains that many of the culture-specific words are not there. For the purpose of marking such words with a proper sense, it is of utmost importance that senses be created for them. The examples for this are:

i. Temple car – The concept here is that of chariots that are used to carry the idols of Hindu gods. The chariot or car is usually used on festival days, when many people pull it.

ii. Auto rickshaws; autos – a motorized version of the traditional rickshaws, has a tin/iron body resting on three small wheels (one in front, two on the rear), a small cabin for the driver (called an auto-wallah in some areas) in the front and seating for three in the rear[17].

The Hindi and Marathi sense markers did not have to resort to hash marking the words from the Tourism text that were not found in the wordnets. Since the two sense inventories are being made indigenously, the lexicographers could inform the wordnet creators about their needs and get the words/senses inserted in the wordnets. The Hindi/Marathi sense markers usually did not find culture-specific words in the sense inventories while sense-marking the tourism text. For example, in Marathi, the word *मंगळागौर (mangalaagaur)* which has the sense

(1) लग्नानंतरच्या श्रावणात मंगळवारी माहेरी व सासरी केलेली पार्वतीची पूजा व समारंभ
- After-marriage Tuesday of ShraavaNa month fathers house or at the house of in-laws performed Parvatis worship
-lagnaantarchyaa shraavaNaat mangalavaarii maaherii va saasarii keleli Parvatichi poojaa va samaarambha
- a religious ceremony to worship Goddess Parvati performed on any Tuesday of the month of Shravan, (fifth month of the Hindu calendar) by a married woman either at her fathers house or at the house of her in-laws.

The word *मंगळागौर* also has the sense of *the name of a goddess* which has not been created so far[18].

Species names:

Though some common flora and fauna names are present in the wordnet, it is lacking in *species names* which commonly occur in the text related to Tourism domain. Names like *elephant grass, Himalayan griffin, blue sheep, snow trout, black-necked grebe, hog deer, Impeyan pheasant, blood pheasant, Bengal florican, etc.* needed to have senses created for them. For example, *snow trout* refers to a particular species of fish which is a cold water riverine and short migratory fish belonging to the family *Cyprinidae* and sub-family *Schizothoracinae*. These are widely distributed in the Himalayan and sub- Himalayan region. This sense would not be captured if the

[17] www.wikipedia.org
[18] http://en.wikipedia.org/wiki/Marathi

components of the name *snow trout* are separated as *snow+ trout*.

In the Hindi wordnet too, for example, the senses of species such as दलदली घड़ियाल (marsh crocodile), or विशाल धनेश (Great Hornbill), and in the Marathi wordnet, the sense of पल्लवपुच्छ कोतवाल (Racket-tailed Drongo) are missing.

Words with affixes

Affixed words have been stored separately in the wordnets. The ones that are not there are such words as *trans*-border, *sub*-alpine, *post*-independence, *eco*-friendly, *multi*-cuisine, *bio*-diverse, stress-*free*, sugar*less*, *etc.*, which appear quite frequently in the documents. It was felt that they, along with their affixes, should be treated as a single word entity and a new sense should be created for them. This was decided keeping in view the fact that an affix may have more than a single sense. For example, the prefix *sub* may have any of the three following senses: (a) under, beneath (as in subterranean, submarine); (b) subsidiary, secondary (as in subplot); and (c) almost, nearly (as in subhuman)[19]. The same would apply to words with suffixes, such as motorable, jeepable, modernish, etc. The suffix *–ish* may mean any of the following, depending on the context in which it is used: (a) Typical or similar to (when appended to many kinds of nouns), as in, *Her face had a girlish charm*; (b) about, approximately (when appended to numbers, especially times and ages), as in, *We arrived at tennish* or *We arrived tennish;* (c) of a nationality, or the language associated with a nationality (when appended to roots denoting names of nations or regions), as in, *Danish, Spanish*, *etc.*; and (d) somewhat (when appended to adjectives), as in, *His face had a greenish tinge*.

In Hindi too, one comes across a suffix like वाला (*wala*) which gives different meanings to the words it is attached to. For example, when it occurs with words like दूध (*doodh* - milk) or मिठाई) mithai - sweets), it conveys the sense of seller of these things. With words like गाड़ी (gaadii – vehicle) or बँगला (bangalaa – bungalow/cottage), it refers to the owner of these. A word like पुलिसवाला (pulisawala - policeman) would mean a person in the police force. When the word is, say, दिल्लीवाला, then this suffix denotes a person belonging to Delhi. Meaning of about to is apparent when the word is आने वाला (aanewala – about to come) or जाने वाला (jaanewala – about to go). An expression like मूँछवाला (moustached – having moustache) gives altogether different meaning because of this suffix *wala*. A different kind of case was encountered in Marathi; for example, the word आपलेपणा (aapalepaNaa - intimacy, closeness) has the suffix पणा (paNaa) which denotes a state, and the whole word conveys the sense of being familiar or close to someone. However, both these words are not found in the respective wordnets for the reason that all the words with the affixes have not been incorporated. The list of such words would be extremely exhaustive and hence a decision has been taken to include them in the later stages of wordnet development.

Multi-words in the corpus

Tourism corpus contains descriptions of places and landscapes, and hence one comes across many rather unexpected multi-words as translation candidates which are not found in the wordnet. Given the frequency of the appearance of compound-word and multi-word expressions, the coverage in the wordnets is insufficient. In the framework of WSD, this becomes an important concern.

There are two kinds of multiword expressions (MWE): one which can have compositional interpretation and the other conveying the non-compositional. Machine cannot infer non compositional multiword expressions, so they have to be stored in the sense repositories. For example, the English expression *green card* conveys the sense of *a card that identifies the bearer as an alien with permanent resident status in the United State.* This could not have come from the meanings of the individual components of this MWE and so we find it stored in the dictionaries, including the wordnet. An example of this in Hindi is *चूल्हा चौका* (cuulhaa-caukaa). In the sentence

(3) इन्हें आगे चलकर चूल्हा चौका ही तो संभालना है
- They in future the work of the kitchen take care of -aux
- inheM aage cala kara cuulhaa-caukaa hii to samhaalanaa hai

[19] http://en.wiktionary.org/wiki/Wiktionary

- They have to take care of only the work of the kitchen in the future,

the multi-word has the sense of रसोई से संबंधित काम (rasoi se sambandhit kaam - *the work related to the kitchen)*. Here, the sense of the components *चूल्हा* (cuulhaa), meaning

(4) *मिट्टी, ईंट या लोहे का बना आग का पात्र जिस पर भोजन पकाते हैं*
- clay, bricks, or iron made artifact on which food cooked is
- mittii, iiMta yaa lohe kaa banaa aag kaa paatra jis para bhojan pakaate hain
- an artifact made of clay, bricks or iron used for cooking food),

and *चौका* (caukaa) meaning *भोजन बनाने का कमरा या स्थान* (bhojan banaane kaa kamaraa yaa sthaan - a room or place for cooking food) would not convey the sense of रसोई से संबंधित काम (rasoi se sambandhit kaam) or the work related to the kitchen.

The MWEs conveying compositional interpretation pose a challenge to the sense markers as they can have multiple senses where some of them are compositional and some are not. The Hindi example of such a case is the expression धूप-छाँव (sunshine and shade). On the literal level, this is a compositional expression, if used in a sentence like the following:

(6) नैनीताल में धूप-छाँव के बीच तालों की सैर का अपना ही मज़ा है।
- Nainital in sunshine and shade amidst lakes' visit own enjoyment-aux
- Nainital mein dhuup-chhanv ke beech taa loM kii sair kaa apnaa hii mazaa hai
- Visiting the lakes in Nainital amidst sun shine and shade has its own enjoyment.

On a metaphorical level, this refers to the ups and downs in one's life; as in *जीवन की धूप-छाँव,* (jeevan kii dhuup-chaanv) which is non-compositional. So is its third usage, as in the expression *धूप-छाँव साड़ी* (dhuup-chhanv saree), where it conveys the sense of a colour which is composed of two intermingling hues.

Furthermore, there is no consistency in the way the compositional expressions appear in the text – they may be written with a hyphen, or without a hyphen with just a blank space in between the components, or they could appear as one unit. Expressions such as *sun-washed* (beaches), *snow-laden* (mountains), *low-impact* (camping) were not present in the wordnet. As a solution it was decided to remove the hyphen and tag the words separately. Thus in the expression *sun-washed* the word *sun* was given the sense – *a typical star that is the source of light and heat for the planets in the solar system*; and *washed* was sense-marked as - *wet as from washing*. In this manner the sense of the entire multiword was tagged. In Hindi and Marathi, the expression *नारी-निकेतन* (naarii-niketan) was dealt with in the same way.

Adjectival Phrases

In the Tourism domain, the sense markers came across a number of adjectival phrases which acted like pre-modifiers of nouns and, at times, behaved like predicatives too. The expressions such as *melt-in-the-mouth* chocolates, a *never-before* adventure, *get-away-from-it-all* appeal, the frenetic *cigarette-and-coffee* pace, etc. have been decided to be lexicalised. Similar is the case with phrases like *sit-out* (as in *to rest in the sit-out*), *a must check-out* (as in *A must check-out is the restaurant at the hotel*).

Figurative language

The sense-markers working in English language came across quite a few idioms and figurative expressions in the tourism corpus as there is a proclivity to use such language in this particular domain. However, this tendency was not found so much in the Hindi and the Marathi corpuses as they comprise of translated matter from English where such examples tend to acquire their literal senses on translation. For English sense-marking, it was decided that these should be lexicalized and thus were enclosed within # marks. For example, in a sentence like *The ambience is good but the food is not much to write home about*, the idiom *not much to write home about* would be lexicalized as *something being mediocre; not as good as you expected or to not be especially good or exciting*. In another sentence, *The palace situated in the lake is a gem of a place for the tourists,* the word *gem* is used metaphorically, conveying the sense of *something that is valued for its beauty or perfection.*

4 Other Challenges

There were some other language issues in the tourism corpus which the sense-markers encoun-

tered. Decisions regarding each category of such words and phrases were taken keeping in mind the general principles of English grammar as well as the requirements of machine translation.

4.1 The participle issue

In words such as *glowing diamonds, shaking towers or fluttering butterflies*, the words *glowing, shaking* and *fluttering* did not have a sense given under the adjective category as here the verb participle is acting as an adjective. Therefore, it was decided to pick up the sense from the verb category. So, for the word *glowing*, the verb sense conveying *emit a steady even light without flames* was picked up.

4.2 Named Entity issue

In the domain of Tourism, it is but natural to come across a large number of named entities as names of destinations and places of interest, such as Coffin Bay or Port Lincoln. It was decided that here the proper Noun part of the name would not to be tagged, but the common noun part would be tagged. So in the above example, the word Coffin was not tagged, whereas the word Bay was given the sense as *an indentation of a shoreline larger than a cove but smaller than a gulf*. Similarly in Hindi, names such as त्रिपोलिया बाजार (Tripoliya Bazaar) and in Marathi टॅमडील तलाव (Tomdil talaava), which are not available in the Hindi and the Marathi wordnets respectively, had only their common noun parts tagged. Thus बाजार (bazaar), which has the sense of

वह स्थान जहाँ तरह-तरह की चीज़ें बिकती हों

(that place where variety of things are sold)

and तलाव (talaava) which has the sense of मोठे तळे (big lake) were tagged to the respective senses.

5 Inadequacies of the English, Hindi and Marathi wordnets as encountered by sense-markers

Besides the above-mentioned challenges, various inadequacies of the wordnets were also encountered. For instance, the adjective senses of many common nouns are absent from the English wordnet; as in the expression *a vegetarian diet,* the word *vegetarian* requires an adjective sense which should mean *consisting primarily or wholly of vegetables and vegetable products*; or the word *budget* as used in the expression *a budget hotel* where the sense of budget should consists of *being as appropriate for a restricted budget or inexpensive*.

Proper nouns such as *Moghul, Rajput,* or *Khmer,* which appear in the wordnets, do not have their adjective senses, as in *the Khmer culture*, or *the Moghul architecture*.

The adverb sense of many words needs to be included in the wordnet. For example: *chillingly* - as in *the worlds most chillingly famous horror attraction*; *fleetingly* - as in *visitors do not merely wanting to experience the attractions fleetingly*; *on-line* - as in *to book accommodation on-line*; *archaeologically* – as in *the cave being archaeologically significant*, *etc.*

Some verbs appear in the wordnet only in the intransitive sense. For example, the verb *dilate* has the sense *become wider* in the sentence "*His pupils were dilated*"; but its transitive sense, that of make wider which should correspond to sentence like *It dilates the blood vessels* has not been given in the English wordnet.

The sense-markers also came across concepts in the Princeton wordnet which, they felt, should have had a broader sense but they appeared in a restricted form. For example, the word *Lord* in the wordnet is defined as *a term referring to the Judeo-Christian God*. When the concept of *Lord* is found in other religions, for example, *Lord Rama* or *Lord Buddha*, it becomes difficult to tag it to the sense quoted above. It would have been much better to have the gloss like a term referring to *God* as given in the English wordnet, instead of restricting it to the Judeo-Christian religion. The words like *Creator, demon, gateway, etc.* posed the same kind of challenge.

6 Comparison with Other Annotation Tasks

Other annotation efforts include Part-of-speech tagging, Chunking and Named Entity Recognition.

	Sense Marking	POS Tagging	NER	Chunking
Options	Typically large	2-3	2-3	Very little
Training Corpus	Very large	30-60K	30K	20K
Complexity of Algorithm	Highly Complex	Medium	Medium	Simple
Language Proficiency	High	Medium	Little	High
Time Taken	Typically much	Not much	Little	Much
Inter Annotator Agreement	Low	Medium	High	High
Language Divergence	Not Affected	Affected	Not Affected	Affected

Table 1: Comparison of Annotation Tasks

Table 1 above shows a comparative study of all the annotation tasks.

Part-of-speech tagging, also called grammatical tagging or word-category disambiguation, is the process of marking up the words in a text (corpus) as corresponding to a particular part of speech, based on both its definition, as well as its context - *i.e.*, relationship with adjacent and related words in a phrase, sentence, or paragraph. Chunking, in computational linguistics, is a method for parsing natural language sentences into partial syntactic structures (noun groups, verbs, verb groups, etc.). Named Entity annotation task seeks to locate and classify atomic elements in text into predefined categories such as the names of persons, organizations, locations, expressions of times, quantities, monetary values, percentages, etc.[20]

7 Conclusion

In this paper we have discussed problems faced in annotating tourism domain corpora in three languages using the sense repositories of English, Hindi and Marathi wordnets. English, as used across the world, varies from country to country and state to state. It is, therefore, a challenge to accurately mark words with their senses. Culture specific words like *curry/gravy* from India, *wanton* from China and *Doro Wat* from African cuisine are such cases in point. Similar is the case with Hindi which has wide variations inside and outside India. Besides culture specific words, different region-specific usages of existing words (meaning expansion), absence of words and absence of senses of existing words are other challenges facing a sense annotator. Of course, the wordnet built for a particular language cannot always accommodate borrowed words, borrowed senses, and meaning contractions and expansions influenced by other languages. But in an increasingly globalized world where code mixing and all other phenomena as listed above are becoming a norm, lexical resources have to evolve strategies of staying useful while maintaining purity and faithfulness to the languages and cultures they represent.

In conclusion, we would like to say we found problems in assigning senses to

1. Culture-specific words
2. Words particular to Indian English
3. Words denoting fractional quantities
4. Species names
5. Words with affixes
6. Multiword expressions
7. Adjectival phrases
8. Figurative language

Partitioning the wordnet into id-regions dedicated to native words and senses, expanded senses and borrowed senses is a concrete suggestion our paper makes to wordnet creators of all languages.

Plans are underway to see that the wordnets link with DBpedia[21], Wikipedia[22] and Yago[23].

References:

Collins Essential English Dictionary 2nd Edition 2006 © HarperCollins Publishers 2004, 2006

Collins Essential Thesaurus 2nd Edition 2006 © HarperCollins Publishers 2005, 2006

Date, Yashwant Rao and Karve, Chintamana Ganesha, 1995. *Maharashtra Shabdakosh.* Varada Books, Pune, Maharashtra

Dhongde, Ramesh, 2009. *Oxford English-Marathi Dictionary*. Oxford University Press, New Delhi

Kernerman English Learners Dictionary © 1986-2008 K Dictionaries Ltd and partners

The American Heritage® Dictionary of the English Language, Fourth Edition copyright ©2000 by Houghton

www.khandbahale.com/englishmarathi

[20] http://en.wikipedia.org/wiki

[21] http://en.wikipedia.org/wiki/DBpedia

[22] http://en.wikipedia.org/wiki

[23] http://www.mpi-inf.mpg.de/yago-naga/yago