Natural Language Processing: A Perspective from Computation in Presence of Ambiguity, Resource Constraint and Multilinguality

Pushpak Bhattacharyya

Professor of Computer Science and Engineering at IIT Bombay

Language is a hallmark of intelligence, and endowing computers with the ability to analyze and generate language- a field of research known as Natural Language Processing (NLP)- has been the dream of Artificial Intelligence (AI). In this paper we give a perspective of NLP from the point of view of ambiguity processing and computing under resource constraint. Language is fraught with ambiguity at all levels, be they morphemes, words, phrases, sentences or paragraphs. We first discuss these ambiguities with examples. Then we take a particular case of disambiguation- word sense disambiguation (WSD)- and discuss its solution in the face of multilinguality and resource constraint, namely, scarcity of annotated data. Multilinguality is one of the powerful instruments of leveraging shared resource.

Index Terms — Ambiguity, Annotated Data, Multilingual Computation, Wordnets, Word Sense Disambiguation

I. Introduction

NATURAL language processing is the task of analyzing and generating by computers, languages that humans speak, read and write [1][2][3].NLP is concerned with questions involving three dimensions: *language, algorithm and problem*. Figure 1 expresses this point. On the language axis are different natural languages and linguistics. The problem axis mentions different NLP tasks like morphology, part of speech tagging etc. The algorithm axis depicts mechanisms like HMM, MEMM, CRF etc. for solving problems.

The goal of natural language analysis is to produce knowledge representation structures like predicate calculus expressions, semantic graphs or frames [4]. This processing makes use of foundational tasks like morphology analysis, Part of Speech Tagging, Named Entity Recognition, both shallow and deep Parsing, Semantics Extraction, Pragmatics and Discourse Processing. The example below illustrates these tasks:

Example 1 : Conversation between a mother and her son:
Mother: *Son, get up quickly. The school is open today. Should you bunk? Father will be angry. Father John complained to your father yesterday. Aren't you afraid of the principal?*
Son: *Mummy, it's a holiday today!*

Processing of the above text involves the following:

1. *Son* in the first sentence has ambiguity of sense. Wordnet 2.1 [5] records two senses of *son: male offspring* (the commonly occurring meaning) and *Son of God, i.e., Jesus.* The first sense applies here.
2. *Get up* is non compositional in the sense that individual meanings of *get* and *up* cannot be composed to decipher the meaning. The combination is a phrasal verb, also called, multiword [6].
3. As a unit, *Get up* is ambiguous, meaning *to wake up* or *to rise*. Finally note that in *get up the ladder* the combination is not a phrasal verb.

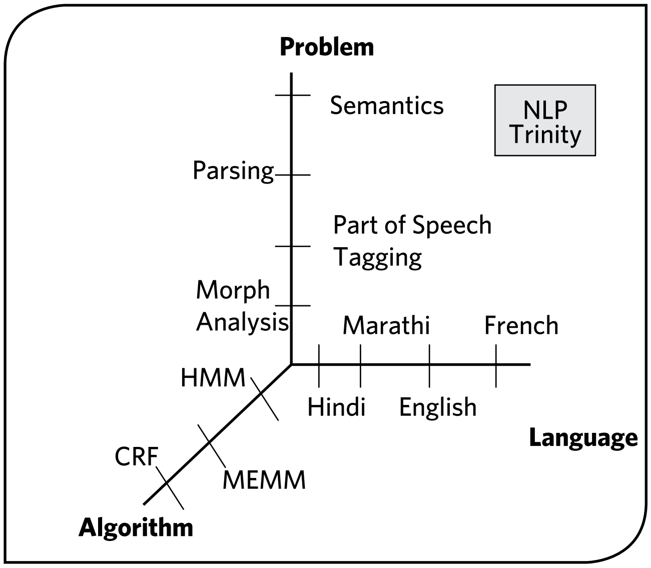

Fig. 1 : *Three dimensions of NLP*

4. *Open* is ambiguous, firstly with respect to part of speech (POS), *i.e.*, noun, verb, adjective or adverb. Here *open* is an adjective, and wordnet records 21 senses of adjective *open*. The 21st sense: *open -- (ready for business; "the stores are open")* applies here.
5. *Should you bunk?* is a sentence with ellipsis [7], *i.e.*, gap or omission which needs to be filled with text from an earlier sentences. Bunk what? Bunk the *school*. Additionally, *Bunk* is ambiguous with respect to both POS and sense.
6. There are three mentions of *father.* The first *father* is most likely the boy's father. This is a matter involving pragmatics. The ambiguity here needs to be resolved with situation specific information, speaker intent *etc.* Understanding *why father will be angry* is a complex inferencing process (not studying leading to bleak future *etc.*).
7. The second *father* is an appellation to *John* who most likely is the principal of the school (again a pragmatics question). The three words *father, John* and *Principal* refer to the same individual. This is the case of co-reference disambiguation. Pronoun to noun binding is a classic case of coreference resolution and is commonly known as anaphora resolution [8]. For example, in the sentence *the dog went near the cat, and it bit it,* it is not clear who bit whom. When the reference to an entity is in the forward direction, instead of the more common backward direction, *e.g.* in *that he will win was clear to Sam*, where *he* is forward bound to *Sam*, the binding is called *cataphora*.
8. *John* in the text is a proper noun. It is important to detect proper nouns in NLP- a problem known as Named Entity Recognition (NER) [9]. In English the NER problem is to an extent ameliorated, since proper nouns are capitalized. In German where all nouns are in capital and in Indian languages where there is no capitalization, separating proper nouns from common nouns is a non-trivial task. An example of this from Hindi is given below:

   Example 2:
   पूजा ने पूजा के लिए फूल तोड़ा
   puujaa ne puujaa ke liye phuul todaa
   puujaa_ERG worship_for flowers plucked
   Puujaa plucked flowers for worship

9. Here the first *puujaa* is the name of a girl and the second *puujaa* means *worship*. Translating this as *worship plucked flowers for worship* is strange, though *puujaa plucked flowers for puujaa* is passable.
10. This is not to say that NER is easy in English. In *Washington voted Washington to power*, first *Washington* is the capital of USA (place name), while the second is George Washington, the first President of USA (person name). Person-place-organization ambiguity in NER is typical in languages.
11. Finally, there are problems of sentencification, tokenization and morphology in any NLP task. These basic tasks too are not free of ambiguity. For example, the verb group *will be going* is indeterminate with respect to the grammatical features of Gender, Number and Person (GNP). The goer could be in 1st/2nd/3rd person (*I, you, he*); the gender could be male or female (English does not mark gender on the verb); the number could be both singular and plural (*I/We, he/they*). Linguistically speaking, English displays high *syncretism, i.e.,* overloading of morphological forms. Hence, even at the level of words, ambiguity needs to be dealt with.

   In what follows, in section II we give an account of the traditionally accepted view of stages of NLP and the associated ambiguities. In section III, we describe a particular kind of ambiguity - the ambiguity of word sense and its solution in a multilingual resource constrained setting. This section also mentions India's large scale activity on multilingual wordnet development, the *Indowordnet* project. Section IV concludes the paper and points to future directions.

II. Stages of NLP and associated ambiguities

Traditionally, NLP - of both spoken and written language- has been regarded as consisting of the following stages:

a. Phonology and Phonetics (processing of sound)
b. Morphology (processing of word forms)
c. Lexicon (Storage of words and associated knowledge)
d. Parsing (Processing of structure)
e. Semantics (Processing of meaning)
f. Pragmatics (Processing of user intention, modeling etc.)
g. Discourse (Processing of connected text)

We describe each stage and the associated ambiguity.

A. Phonology and Phonetics

At this stage utterances are processed. Apart from many challenges due to noise, two common problems are *homophony* and *word boundary recognition*.

Homophony arises when two words sound the same, though their meanings are widely different, e.g., *bank* (embankment of a water body) and *bank* (an institution where financial transactions are held). Homophony, it is surmised, originates in borrowing and adapting of words from a foreign language (*e.g.,bank* in the financial sense in English came from *banque* in German).

Near homophony is more common and causes difficulty, especially in rapid speech. e.g. *fox* and *folks.* We normally do not falter in understanding, because syntactic clues, context and world knowledge comes to the rescue.

Word boundary detection, similarly, is a challenge in case of rapid speech. Consider the following example in Hindi:

Example 3:

आजायेंगे

(आज आयेंगे will come today or आ जायेंगे will come)

aajaayenge

(aaj aayenge or aa jaayenge)

The string can be broken in two ways as shown above with two completely different meanings.

A similar example in English is *I got up late vs. I got a plate*, both of which sound very much the same.

B. Morphology

Words form from root words or lexemes through processes of inflexion, derivation, back formation, clitics and portmanteauing [7]. Languages differ in their morphological richness. Dravidian languages, Turkish, Hungarian and Slavic languages are examples of morphologically rich languages. Chinese and English are examples of relatively simpler morphology.

The main ambiguity at the level of morphology arises from choices available in breaking the word into stem and suffix as well as from choices of features. Here is an example from Marathi:

Example 4:

तो जाई पर्यंत मी राहील
to jaaii paryanta mii raahiil

He going till I will stay

I will stay till he goes

and

मधूप जाई पर्यंत जाऊन परत आला
madhuup jaaii paryanta jaauun parat aalaa

Honeybee jaaii (a flower) till having-gone came back

The honey bee went till the Jaaii flower and came back

The word *jaaii* can be broken into two morphemes *jaa* and *ii* or can be left unbroken, giving two different morphemes.

The problem of feature ambiguity has been illustrated in the section I point no. 9.

New words introduced through technological changes pose challenges for morphology. Following words are new in English:

Justify (as in justifying the right margin)
Xerox (as a verb)
Discomgooglation (discomfort at not being able to use Google)
Communifaking (pretending to talk on the mobile)
Nomophobia (phobia from no mobile)

C. Lexicon

Words are stored in the lexicon with a variety of information that facilitates the further stages of NLP, like question answering, information extraction *etc*. For example, the word *dog* might be stored in the lexicon with information like:

POS (Noun)
Semantic Tag (Animate, 4-legged)
Morphology (takes 's' in plural)

Words typically have multiple meanings even in the same part of speech. *Dog*, for example, means an animal and a very detestable person.

In case of word sense ambiguity two situations are distinguished- *homography* and *polysemy*. Homography like homophony results from foreign word borrowing. Two words are homographic if they are spelt the same, though their meanings are different. Again, *bank* is an example of homography. Polysemy, on the other hand, implies shades of meaning, e.g., *falling of tree* and *falling of a kingdom*.

Word sense or lexical disambiguation refers to the identification of the meaning of an ambiguous word from clues in the context. For example, in *I will withdraw some money from the bank*, the most likely sense of *bank* is the financial institution sense. The senses typically come from a sense repository like the wordnet [5].

The predominant approach in WSD is supervised learning, where a machine is trained with sense annotated corpora [10] [11]. But sense annotated corpora are a costly resource. Our research shows that we can leverage multilinguality and multilingual linked wordnets to reduce demand on annotated corpora. This is the topic of discussion in the next section.

D. Parsing

Parsing or syntactic processing refers to uncovering the hierarchical structure behind a linear sequence of words. For example, the noun phrase (NP) *flight from Mumbai to Delhi via Jaipur on Air India* has the following structure:

[NP_4

[NP_3

[NP_2

[NP_1 [NN flight]]

[PP_1 [P from][NP [NNP Mumbai]]]

]

[PP_2 [P to] [NP [NNP Delhi]]]

]

[PP_3 [P via][NP [NNP Jaipur]]]

]

[PP_4 [P on][NP [NNP Air-India]]]

]

The above is called a bracketed structure after the name given in the Penn Treebank project[1] to the parsed tree data. The above structure shows that *flight* is a noun (NN) which is modified by the attached preposition phrase (PP_1) *from Mumbai* to form NP_1. NP_1 is modified by the attached PP_2 *to Delhi* to form NP_2. NP_2 is modified by the attached PP_3 *via Nagpur* to form NP_3 which in turn is modified by the attached PP_4 *on Air-India* to form NP_4.

Such bracketed structures are created by a Grammar of the language. In the above example, *PP→P NP* is a grammar rule expressing the fact that a preposition phrase is composed of a preposition and a noun phrase.

Now, parsing too faces the challenge of ambiguity called *structural ambiguity*. Structural ambiguity is of two kinds: *scope ambiguity* and *attachment ambiguity*. We give examples of these kinds of ambiguity:

Example 5 (scope ambiguity):

Old men and women were taken to safe locations.

The scope of the adjective (*i.e.*, the amount of text it qualifies) is ambiguous. That is, is the structure *(old men and women)* or *((old men) and women)*?

Another example of scope ambiguity is:

[1] http://www.cis.upenn.edu/~treebank/

Example 6 (scope):

No smoking areas will allow hookahs inside.

Here *no* can qualify the rest of the sentence, meaning thereby *there isn't a smoking area that will allow hookas inside.*

Or

It can qualify only the phrase *smoking areas*, meaning thereby *there are areas designated as no-smoking-areas which, however, allow hookas inside.*

The two meanings are sort of opposite of each other.

Attachment ambiguity arises from uncertainty of attaching a phrase or clause to a part of a sentence. Here are some examples:

Example 7 (attachment ambiguity):

I saw the boy with a telescope.

It is not clear who has the telescope, *I* or *the boy*? In the former case, we say, the preposition phrase *with a telescope* attaches with the verb *saw* with the instrumental case. In the latter the PP attaches to *the boy* as a modifier.

PP-attachment is a classical problem in NLP [12]. The general problem can be stated as follows:

Given the structure

$V\text{-}NP_1\text{-}P\text{-}NP_2$

Where does NP2 attach, V or NP1?

The problem is attempted to be solved by both rule based and machine learning based approaches. In the former, the properties of *V, head (NP1)* and *head (NP2)* are used to formulate rules for attachment. For example, in *I saw the boy with a pony tail*, the PP attaches to *the boy*, since pony tail does not possess the property of instrumentality and *saw with pony tail* does not make sense. Such properties- called selectional preferences come from lexical knowledge bases like wordnet [5] and verbnet[2]. Formulating such rules for deciding attachment is human labour intensive, and ensuring correctness and completeness of the rule base is well nigh impossible.

The alternative approach to solving the attachment problem is machine learning (ML) based. Here one creates annotated corpora of the form:

See the boy with pony tail: V

See the tiger with telescope: N

and then try to teach a machine the conditions for the two kinds of attachment. ML algorithms from simple (Decision Trees [13]) to very complex ones (Graphical Models [14]) have been employed.

It must be clear that ambiguity of attachment arises from the dual role of prepositions, *viz.,* assigning case to nouns with respect to a verb and for modifying a noun phrase. In example 7, *with* can assign instrument case to *telescope* or specify a particular boy *having* a telescope.

Indian languages have *post positions* instead of prepositions, that is, entities that assign case roles follow the noun, and do not precede.

Example 8 (in Hindi):

दूरबीन से लड़के को देखा
duurbiin se ladke ko dekhaa
telescope_with boy_ACC saw
saw the boy with a telescope

The postposition *se* assigns case role to *duurbiin* and follows it.

Attachment ambiguity of the type pp-attachment is not so common in Indian languages which are as a rule SOV (subject-object-verb) languages. Postpositions follow this pattern:

NP1 P NP2 V

(In example 8, *NP1= duurbin, P=se, NP2= ladke, V=dekhaa*)

Postpositions typically assign case and hardly modify the following NP. One exception to this is the genitive case (*of;* Hindi का के की *ka, ke, kii*). But the genitive case marker always links two NPs.

Attachment ambiguity arises with phrases and clauses too, as illustrated below:

Example 9 (attachment ambiguity; phrase):

I saw a tiger running across the field

Who was running: *I* or *the tiger*? The attachment of the phrase *running across the field* is ambiguous.

Example 10 (attachment ambiguity; clause):

I told the child that$_1$ I liked that$_2$ he came to the playground early.

The sentence has two meanings: (a) *I told the child the FACT that I liked his coming early to the ground* and (b) *I told the child WHOM I liked that he came early to the ground.* The ambiguity here comes from the dual role of that, viz., relative pronoun or complementizer. In the former situation, *that$_1$* attaches to child and in the latter situation, the *that$_1$* attaches to *told*.

What all these discussions show is that syntactic ambiguity is as widespread as lexical ambiguity, but is harder to detect. This is the reason one finds far less work reported on this type of ambiguity. This is true of both *constituency parsing* that identifies phrases in a sentence, and *dependency parsing* that finds heads and modifiers [1].

E. Semantics Processing

After word forms and structure have been detected, sentence processing devotes itself to meaning extraction. While the meaning of *meaning* is debatable, there is a general agreement that at the stage of semantic processing, the sentence needs to be represented in one of the unambiguous forms like predicate calculus, semantic net, frame, conceptual dependency, conceptual structure etc. [4]. We at IIT Bombay have for long used the Universal Networking Language (UNL) framework [15]. Figure 1 below illustrates the UNL representation:

[2] http://verbs.colorado.edu/~mpalmer/projects/verbnet

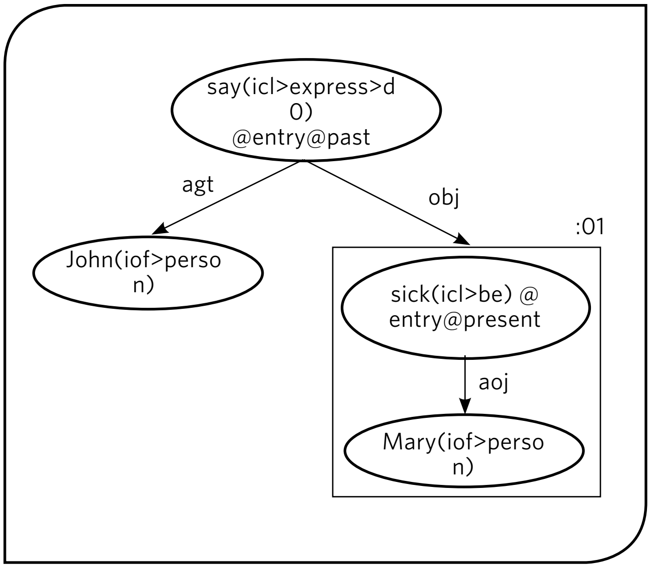

Fig. 2 : *UNL graph of the sentence John said Mary is sick.*

UNL graph is constructed around the main verb of the sentence. The main verb is the *entry* point of the knowledge content of the sentence. Nodes are the concepts- called *Universal Words*, and the edges are semantic relations between concepts. *Attributes* are added to concept nodes to express tense, plurality, main predicate, emphasis, topicalization, speech act etc. Universal words (UWS) are disambiguated entities as indicated by *restrictions* in parentheses attached to them. Thus the building blocks of UNL representation are:

1. UWs with restrictions on them
2. Relations
3. Attributes

In the example in figure 2, *say* is the main predicate and is in the sense of *expressing* something. The relations emerging from the *say* node are *agt* (agent; John is the agent of the say activity) and *obj* (object; what John said). The *obj* arc points to a box containing a subgraph standing for an embedded sentence or clause. Such nodes are called *scope nodes* or *compound UWs* and are typically given an *id* (:01 in figure 2).

Now, semantics extraction faces all the challenges arising out of ambiguities of semantic roles or relations. Some classic examples are:

Example 11:

Visiting aunts can be trying

Here, are the *aunts* visitors (agent role) or are they being visited (object role)? All languages exhibit this kind of ambiguity, as in Hindi:

Example 12:

आपको मुझे मीठाई खीलानी पड़ेगी

aapko mujhe mithaai khilaanii padegii

You_DATIVE by_me sweets fed_OBLIGATION

Or

By_you I_DATIVE sweets fed_OBLIGATION

You will have to feed me sweets

Or

I will have to feed you sweets

Both *I* and *you* have semantic role ambiguity *(agent vs. beneficiary).*

F. Pragmatics Processing

This is one of the hardest problems of NLP and has seen very little progress. The problem involves processing user intention, sentiment, belief world, modals *etc.*- all of which are highly complex tasks. The following humorous exchange illustrates the nature of the problem:

Example 12:

Tourist (checking out of the hotel): Waiter, go upstairs to my room and see if my sandals are there; do not be late; I have to catch the train in 15 minutes.

Waiter (running upstairs and coming back panting): Yes sir, they are there.

Clearly, the waiter is falling short of the expectation of the tourist, since he does not understand the pragmatics of the situation. But *are my sandals there* is an ambiguous question if *user intent* and the situation specificity are considered. This may be either a request for information or a request for action. Larger context, history, intent, sentiment, tone *etc.*- all these come into play, making the task enormously difficult.

G. Discourse Processing

This is the task of processing connected sentences. Section I on introduction brought out the difficulties of the task through a mother-son conversation situation. All the NLP problems discussed so far, surface when we process connected text. In a speaker-listener scenario, the listener continuously produces hypotheses in his mind and updates them about the world the conversation proposes to create, as the following series of sequence of sentences illustrates:

Example 13:

Sentence-1: John was coming dejected from the school

(who is John: most likely a student?)

Sentence-2: He could not control the class

(who is John now? Most likely the teacher?)

Sentence-3: Teacher should not have made him responsible

(who is John now? Most likely a student again, albeit a special student- the monitor?)

Sentence-4: After all he is just a janitor

(all previous hypotheses are thrown away!).

This is the nature of discourse processing. In addition to ellipsis, coreference, sense and structure disambiguation and so on, an incremental building up of the shared world has to carried out.

H. Textual Humour and Ambiguity

We end this section with a glimpse of how ambiguity is at the heart of humour, especially of the textual kind. Computational humour[3] is an actively researched area. Theories of humour suggest that one of the reasons for humour arising is incongruity of views - the so called *incongruity theory*. Incongruity arises from ambiguity, as the examples below show:

Example 14 (Humour and lexical ambiguity):

A car owner after coming back from a party finds the sticker "parking fine" on his car. He goes and thanks the policeman for appreciating his parking skill.

The ambiguity of the word *fine* (*nice vs. penalty*) and the two different meanings picked by the car owner and the policeman give rise to the humour.

Example 15 (Humour and structural ambiguity):

Teacher: What do you think is the capital of Morocco?

Student: What do you think?

Teacher (Angrily): I do not think, I know.

Student: I ... do not think I know.

The attachment ambiguity of *I know* (standalone sentence vs. getting attached to *think*) and the two different attachments picked by the teacher and the student give rise to humour.

III. Resource constrained word sense disambiguation

Having described different kinds of ambiguity in NLP, we now turn to a specific ambiguity called lexical or sense ambiguity. The problem is defined as follows:

Definition (WSD): *Given a text environment W_1, W_2, W_3,… ,W_T,… W_n and a target word W_T in that environment, mark the correct sense id on W_T based on the clues in the environment. The id will come from a lexical resource like the wordnet.*

We have mentioned in section II.C that processing of lexical or sense ambiguity plays a critical role in NLP. The area of word sense disambiguation (WSD) is well investigated, with all four approaches, *viz., knowledge based, supervised, semi-supervised* and *unsupervised* having been tried [10][11].

The predominant approach to WSD, however, has been machine learning based. Annotated corpora in large amount are used to train a machine to perform WSD. However, creation of sense marked corpora is an expensive proposition requiring large investment in the form of expert manpower, time and money. The list below gives an idea about training data for WSD:

- *SemCor*: ~200000 sense marked words for English in general domain [16]
- Domain specific multilingual Sense marked corpora created at IIT Bombay[3]
  - English: Tourism (~170000), Health (~150000)
  - Hindi: Tourism (~170000), Health (~80000)
  - Marathi: Tourism (~120000), Health (~50000)
- 12 man years for each <L,D> combination

WSD in general domain has not been able to achieve very high accuracy. On the other hand, domain specific WSD has been quite successful. Our long standing work on domain specific multilingual WSD gives us a vision of WSD research expressed in the following matrix (Table 1).

Table 1: domain vs. language matrix for WSD

		Languages							
		Hindi	**Marathi**	**Tamil**	**Telugu**	..	..	..	**Kannada**
Domain	**Tourism**	X				..	..	..	
	Health		X			..	..	..	
	Finance					..	..	..	
	Sports					..	..	..	
	..	..	..	..	..	..	..	..	..
	..	..	..	..	..	..	..	..	..
	Politics					..	..	..	

We hope that *if we can create a WSD system for a particular language-domain pair, we can extend the system along both the language axis and the domain axis, thereby filling the whole matrix and creating general purpose all words WSD system.*

The key idea here is to use *projection* which we explain in next few subsections. The discussion is based on our work on multilingual, resource constrained WSD [17]-[21], which gives hope for the above vision.

A. Parameters for WSD

Consider the following sentence:

The river flows through this region to meet the sea.

The word *sea* is ambiguous and has three senses as given in the Princeton Wordnet (PWN):

S1: (n) sea (a division of an ocean or a large body of salt water partially enclosed by land)

S2: (n) ocean, sea (anything apparently limitless in quantity or volume)

S3: (n) sea (turbulent water with swells of considerable size) "heavy seas"

The first parameter for WSD is obtained from *Domain specific sense distributions.* In the above example, the first sense is more frequent in the tourism domain (verified from manually sense marked tourism corpora).

There are other parameters for WSD as follows [19]:

Wordnet-dependent parameters

belongingness-to-dominant-concept

conceptual-distance

semantic-distance

Corpus-dependent parameters

corpus co-occurrences.

However, we find from our study and systematic procedures like the ablation test that domain specific sense distribution information is the most important parameter for WSD.

B. Scoring Function for WSD

Based on the above parameters, we desired a scoring function which:

(1) Uses the strong clues for disambiguation provided by the

[3] http://en.wikipedia.org/wiki/Computational_humor

monosemous words and also the already disambiguated words.

(2) Uses sense distributions learnt from a sense tagged corpus.

(3) Captures the effect of dominant concepts within a domain.

(4) Captures the interaction of a candidate synset with other synsets in the sentence.

We have been motivated by the Energy expression in Hopfield network [22] in formulating a scoring function for ranking the senses. The correspondences are as follows:

Neuron → Synset

Self-activation → Corpus Sense Distribution

Weight of connection between two neurons → Weight as a function of corpus co-occurrence and Wordnet distance measures between synsets

$$S* = \arg\max_{i}(\theta_i \times V_i + \sum_{j \in J} W_j \times V_i \times U_i) \qquad (1)$$

where,

S=best possible sense*

J=set of disambiguated words

θ_i=*BelonginessToDominantConcept(*S_i*)*

V_i=*P(*S_i*|W)*

Uj=P(sense assigned to $W_j|W_j$*)*

W_{ij}=*CorpusCooccurence(*S_i*,*S_j*) X*

1/WNConceptualDistance(S_i,S_j) X

1/WNSemanticGraphDistance(S_i,S_j)

The component θ_i*V_i is the energy due to the self activation of a neuron and can be compared to the corpus specific sense of a word in a domain. The other component W_{ij}*V_i*U_j coming from the interaction of activations can be compared to the score of a sense due to its interaction in the form of corpus co-occurrence, conceptual distance, and wordnet-based semantic distance with other words in the sentence. The first component thus captures the rather *static corpus sense*, whereas the second expression brings in the *sentential context*.

C. WSD Algorithm employing the scoring function

We give a greedy iterative algorithm IWSD as follows:

Algorithm 1: *performIterativeWSD(sentence)*

1. Tag all monosemous words in the sentence.
2. Iteratively disambiguate the remaining words in the sentence in increasing order of their degree of polysemy.
3. At each stage select that sense for a word which maximizes the score given by Equation (1)

For evaluation of the algorithm we used sense marked corpora in the tourism domain. Prior to our work large scale all words domain specific corpora were not available in any language including English. Hence, as part of our earlier work, we set upon the task of collecting data from two domains, *viz., Tourism* and *Health* for *English*. The data for Tourism domain was downloaded from Indian Tourism websites whereas the data for Health domain was obtained from two doctors. The data was then sense annotated by two lexicographers adept in English. Princeton Wordnet 2.1[4] was used as the sense inventory. Some files were sense marked by both the lexicographers, and the Inter Tagger Agreement (ITA) calculated from these files was around 85%.

This was a first of its kind effort at collecting all-words domain specific sense marked corpora. This data is now available freely for research purposes[5] and should help to advance the research for domain-specific all-words WSD. Tables 2, 3 give statistics about the data.

Table 2 : Number of polysemous words per category in each domain.

	English		
	Tourism	**Health**	**SemCor**
Noun	62636	53173	66194
Verb	30269	31382	84815
Adjective	25295	21091	24946
Adverb	7018	6421	11803
All	125218	112067	187758

Table 3: Average degree of wordnet polysemy of polysemous words per category in each domain

	English		
	Tourism	**Health**	**SemCor**
Noun	3.74	3.97	3.55
Verb	5.01	5.31	4.28
Adjective	3.47	3.57	3.26
Adverb	2.89	2.96	2.72
All	3.93	4.15	3.64

Table 4 : Average degree of corpus polysemy of polysemous words per category in each domain

	English		
	Tourism	**Health**	**SemCor**
Noun	1.68	1.57	1.90
Verb	2.06	1.99	2.44
Adjective	1.67	1.57	1.70
Adverb	1.81	1.75	1.79
All	1.77	1.68	1.99

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

[5] http://www.cfilt.iitb.ac.in/wsd/annotated_corpus

We compared our algorithm with the following state of the art algorithms:

i. **IWSD:** our iterative WSD algorithm

ii. **EGS:** exhaustive graph search algorithm using our scoring function

iii. **PPR:** a state of the art knowledge based approach

iv. **SVM:** a state of the art SVM based supervised approach reimplemented by us

v. **McCarthy *et al.*, [23]:** a state of the art unsupervised approach, reimplemented by us

vi. **RB:** randomly selects one of the senses of the word from the Wordnet

vii. **WFS:** assigns the first sense of the word from the Wordnet.

viii. **MFS:** assigns the most frequent sense of the word as obtained from an annotated corpus

The first two results (IWSD and EGS) give an idea about the performance of the proposed scoring function in two different settings (greedy v/s iterative). The next three results (PPR, SVM and [23]) provide a comparison with other state of the art algorithms for WSD. The final three results provide a comparison with typically reported baselines for WSD.

Table 5 : Precision, Recall and F-scores of different WSD algorithms.

Algorithms	Tourism			Health		
	P%	R%	F%	P%	R%	F%
SVM	78.82	78.76	78.79	79.64	79.59	79.61
IWSD	77.00	76.66	76.83	78.78	78.42	78.60
MFS	77.60	75.20	76.38	79.43	76.98	78.19
WFS	62.15	62.15	62.15	64.67	64.67	64.67
PPR	53.1	53.1	53.1	51.1	51.1	51.1
[McCarthy et al., 2007)	51.85	49.32	50.55	—	—	—
RB	25.50	25.50	25.50	24.61	24.61	24.61

Interesting observations here are that our algorithm (IWSD) beats the random baseline (RB) and wordnet first sense (WFS) approaches by a large margin, which lends credence to the approach. Most frequent sense (MFS) from the corpus is very difficult to beat, though this statistics is usually not available, requiring as it does large amounts of sense marked corpora. Our approach comes very close to the MFS value. Among other approaches only SVM beats IWSD.

D. Parameter Projection

Now that the efficacy of IWSD has been established, we discuss what to do about its resource requirement. The scoring function shows that we need two resources for the algorithm: wordnet and sense marked corpora, both of which are costly resources.

This situation leads to the following important question: can the effort required in constructing multiple wordnets and collecting sense marked corpora in multiple languages be avoided? Our findings suggest that the cost and time needed for developing wordnets in multiple languages can be reduced by using the expansion approach [24]. Our efforts at reducing the annotation cost in multiple languages are also centered on such a novel synset based multilingual dictionary where the synsets of different languages are aligned and thereafter the words within the synsets are manually cross-linked [25].

Table 6: Synset based multilingual dictionary: *MultiDict* [25]

Concepts	L1 (English)	L2 (Hindi)	L3 (Marathi)
04321: a youthful male person	(Malechild, boy)	[लड़का (ladkaa), (बालक् (baalak), (बच्चा (bachchaa)]	[मुलगा (mulgaa), (पोरगा (porgaa), (पोर (por)]

Table 6 shows the structure of MultiDict. The concept of *boy* has the Princeton wordnet 2.1 id of 04321 and is linked to synsets of 18 Indian languages only two of which- *Hindi* and *Marathi*- are shown in the table. Actually Hindi serves as the pivot language with other Indian languages linking to it as part of the Indowordnet effort [26].

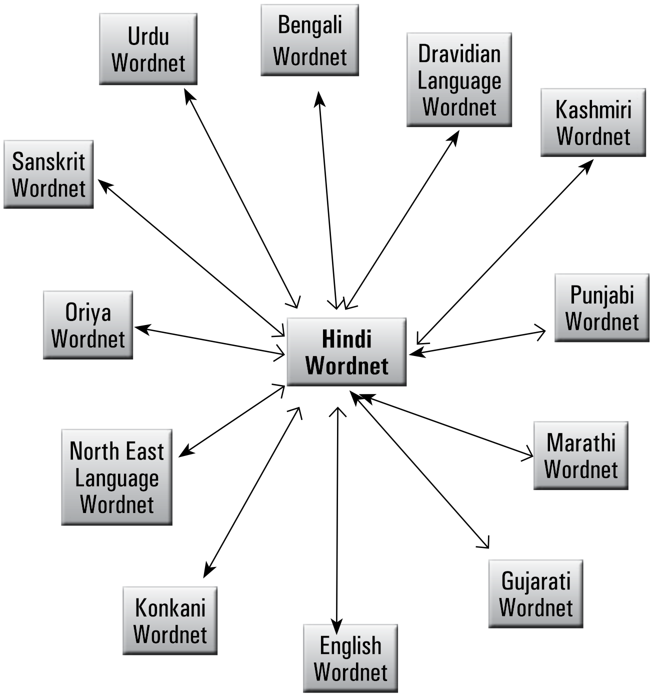

***Fig. 3 :** Indowordnet: linked structure of Indian language wordnets [26]*

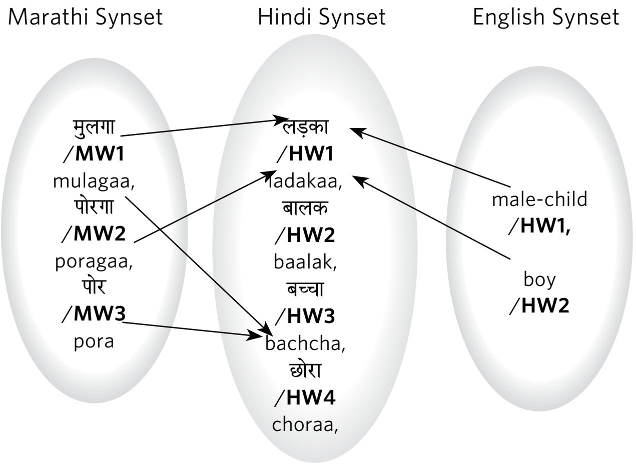

Fig. 4 : cross linkages within linked synsets

Suppose a word (say, *W*) in language L_1 (say, Marathi) has ***k*** senses. For each of these ***k*** senses we are interested in finding the parameter ***P(S***$_i$***|W)-*** which is the probability of sense ***S***$_i$ given the word ***W*** expressed as:

$$P(S_i \mid W) = \frac{\#(S_i, W)}{\sum_j (S_j, W)} \quad (2)$$

where '#' indicates 'count-of'. Consider the example of two senses of the Marathi word सागर *{saagar}, viz., sea* and *abundance* and the corresponding cross-linked words in Hindi (Fig. 5 below):

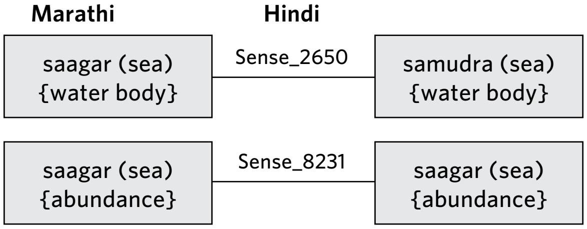

Fig. 5 : *Two senses of the Marathi word सागर (saagar), viz., {water body} and {abundance}, and the corresponding cross-linked words in Hindi*[6].

The probability *P({water body}|saagar)* for Marathi is

$$\frac{\#(\{Water\_body\}, saagar)}{\#(\{Water\_body\}, saagar) + \#(\{abundance\}, saagar)}$$

We propose that this can be approximated by the counts from Hindi sense marked corpora by replacing *saagar* with the cross linked Hindi words *samudra* and *saagar*, as per Fig. 5:

$$\frac{\#(\{Water\_body\}, samudra)}{\#(\{Water\_body\}, samudra) + \#(\{abundance\}, saagar)}$$

Thus, the following formula is used for calculating the sense distributions of Marathi words using the sense marked Hindi corpus from the same domain:

$$P(S_i \mid W) = \frac{\#(S_i, cross\_linked\_hindi\_words)}{\sum_j \#(S_j, cross\_linked\_hindi\_words)}$$

Note that we are not interested in the *exact* sense distribution of the words, but only in their relative values.

To prove that the projected relative distribution is faithful to the actual relative distribution of senses, we obtained the sense distribution statistics of a set of Marathi words from a sense tagged Marathi corpus (we call the sense marked corpora of a language its *self corpora*). These sense distribution statistics were compared with the statistics for these same words obtained by *projecting from* a sense tagged Hindi corpus. The results are summarized in table 6.

The third row of table 6 shows that whenever ठिकाण *(thikaan) (place, home)* appears in the Marathi tourism corpus there is a much higher chance of it appearing in the sense of *place* (96.2%) then in the sense of *home* (3.7%). Column 5 shows that the relative probabilities of the two senses remain the same even when using projections from Hindi tourism corpus (i.e. by using the corresponding cross-linked words in Hindi).

Table 6: Comparison of the sense distributions of some Marathi words learnt from Marathi sense tagged corpus with those projected from Hindi sense tagged corpus

Sr. No	Marathi Word	Synset	P(S|word) as learnt from sense tagged Marathi corpus	P(S|word) as projected from sense tagged Hindi corpus
1	किंमत (kimat)	{ worth }	0.684	0.714
		{ price }	0.315	0.285
2	रस्ता (rasta)	{ roadway }	0.164	0.209
		{road, route}	0.835	0.770
3	ठिकाण (thikan)	{ land site, place}	0.962	0.878
		{ home }	0.037	0.12
		{abundance}	0	0

The other corpus based parameter *corpus cooccurrence* was similarly projected from Hindi to Marathi and it was found that the distribution remains faithful to the original distribution.

IWSD was run on Marathi and Bengali test corpora being trained on Hindi training corpora. That is IWSD parameters were learnt from Hindi and *used* for Marathi and Bengali.

Table 7 shows the accuracy figures with and without projection. The values lend ample credence to the idea of projection from one language to another. The performance using projection falls below the performance, when trained on

[6] Sense_8231 shows the same word saagar for both Marathi and Hindi. This is not uncommon, since Marathi and Hindi are sister languages.

own language data by about 10%, but is above the wordnet baseline by about 20%. The behaviour repeats even when a familially distant language, *viz.,* Tamil is chosen.

Table 7: Precision, Recall and F-scores of IWSD, PageRank and Wordnet Baseline. Values are reported with and without parameter projection.

Algorithm	Language					
	Marathi			Bengali		
	P %	R %	F %	P %	R %	F %
IWSD (training on self corpora; no parameter projection)	81.29	80.42	80.85	81.62	78.75	79.94
IWSD (training on Hindi and reusing parameters for another language)	**73.45**	**70.33**	**71.86**	**79.83**	**79.65**	**79.79**
PageRank (training on self corpora; no parameter projection)	79.61	79.61	79.61	76.41	76.41	76.41
PageRank (training on Hindi and reusing parameters for another language)	**71.11**	**71.11**	**71.11**	**75.05**	**75.05**	**75.05**
Wordnet Baseline	58.07	58.07	58.07	52.25	52.25	52.25

IV. Conclusion

In this paper we have discussed Natural Language Processing from the perspective of ambiguity, multilinguality and resource constraint. First we described different kinds of ambiguity that obtain in NLP starting from the lowest level of processing, viz., morphology to the highest level, viz., pragmatics and discourse. Then we took up one specific ambiguity, viz., word sense and described ways of tackling it under constraints of resource, *viz.*, annotated corpora. Multilinguality was leveraged in the sense of projecting sense distributions in the corpora from one language to another and wordnet parameters like distance between senses. Performance with and without projection were compared, and the idea of projection seemed well founded.

Future work consists in exploring limits of multilinguality. Domain specific sense distributions tend to be universal across languages. This is a highly redeeming observation in the face of resource scarcity. A crucial instrument to leverage multilinguality is linked lexical resources like wordnets. This underlines the importance of investing in multilingual knowledge networks.

Acknowledgment

The author would like to acknowledge the TDIL (Technology Development in Indian Languages) program of Department of Information Technology, Ministry of Communication and Information Technology, Govt. of India, for supporting this research. I am also thankful to my students and colleagues- too numerous to name- interaction with whom has enriched my foundation and perspective.

References

[1] Daniel Jurafsky and James H. Martin. *Speech and Langugage Processing*. Prentice Hall, New Jersey, second edition edition, 2008.

[2] Christopher Manning and Hinrich Schutze. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA, 1999.

[3] James Allen, *Natural Language Understanding*. Pearson Education, 1995.

[4] R. J. Brachman and H. Levesque, *Readings in Knowledge Representation*, Morgan Kaufmann, 1985.

[5] Christiane Fellbaum, editor. *WordNet: An Electronic Lexical Database*. MIT Press, 1998.

[6] I Sag, T Baldwin, F Bond, A Copestake, D Flickinger, "Multiword Expressions a Pain in the Neck for NLP", in *Computational Linguistics and Intelligent Text Processing*, pp. 189-206, 2002**.**

[7] A. Akmajian, R.A. Demers, A.K. Farmer, R.M. Harnish, *Linguistics an Introduction to Language and Communication*, Prentice Hall, 1995.

[8] Ruslan Mitkov. *Anaphora Resolution*. Longman, 2002.

[9] Daniel M. Bikel, Richard Schwartz and Ralph M. Weischedel, "An Algorithm that Learns What's in a Name", in *Machine Learning Journal Special Issue on Natural Language Learning*, 1999.

[10] Eneko Agirre and Philip Edmonds. *Word Sense Disambiguation: Algorithms and Applications*. Springer, 2007.

[11] R. Navigli, "Word Sense Disambiguation: a Survey", *ACM Computing Surveys*, 41(2), ACM Press, 2009.

[12] Adwait Ratnaparkhi, Jeff Reynar, and Salim Roukos, "A Maximum Entropy Model for Prepositional Phrase Attachment", *In Proceedings of the ARPA Human Language Technology Workshop*, 1994.

[13] Tom Mitchell, *Machine Learning,* McGraw Hill, 1997.

[14] Daphne Koller and Nir Friedman, *Probabilistic Graphical Models*, MIT Press, 2009.

[15] H. Uchida,M. Zhu, T. Della Senta, *The UNL, A Gift for a Millennium, UNU/IAS Press, 1999..*

[16] Miller, George A., Claudia Leacock, Randee Tengi, and Ross T. Bunker, "A semantic concordance", *In Proceedings of the workshop on Human Language Technology*, 1993.

[17] Mitesh Khapra, Sapan Shah, Piyush Kedia and Pushpak Bhattacharyya, "Projecting Parameters for Multilingual Word Sense Disambiguation", *Empirical Methods in Natural Language Prfocessing (EMNLP09),* Singapore, August, 2009.

[18] Mitesh Khapra, Saurabh Sohoney, Anup Kulkarni, and Pushpak Bhattacharyya, "Value for Money: Balancing Annotation Effort, Lexicon Building and Accuracy for Multilingual WSD", *In Proceedings of the 23rd International Conference on Computational Linguistics*, Beijing, 2010.

[19] Khapra, Mitesh, Sapan Shah, Piyush Kedia, and Pushpak Bhattacharyya, "Domain Specific Word Sense Disambiguation Combining Corpus Based and Wordnet Based Parameters", *In 5th International Conference on Global Wordnet*, Mumbai, 2010.

[20] Mitesh Khapra, Salil Joshi, Arindam Chatterjee, and Pushpak Bhattacharyya, "Together We can: Bilingual Bootstrapping for WSD", *In Annual Meeting of the Association of Computational Linguistics (ACL 2011),* Oregon, USA, 2011.

[21] Mitesh Khapra, Salil Joshi and Pushpak Bhattacharyya, "Help Me and I will Help You: A Bilingual Unsupervised Approach for Estimating Sense Distributions using Expectation Maximization", *in 5th International Joint Conference on Natural Language Processing (IJCNLP*

2011), Chiang Mai, Thailand, November 2011.

[22] J. J. Hopfield, "Neural Networks and Physical Systems with Emergent Collective Computational Abilities", *In Proceedings of the National Academy of Sciences of the USA,* 79(8), 1982.

[23] Diana McCarthy, Rob Koeling, Julie Weeds, and John Carroll, "Unsupervised Acquisition of Predominant Word Senses", *Computational Linguistics*, 33(4), pp. 553–590, 2007.

[24] Piek Vossen, "EuroWordNet: a Multilingual Database of Autonomous and Language-Specific Wordnets Connected via an Inter-Lingual-Index" In: *special issue on multilingual databases. International Journal of Linguistics 17/2* , 2004

[25] Rajat Mohanty, Pushpak Bhattacharyya, Prabhakar Pande, Shraddha Kalele, Mitesh Khapra, and Aditya Sharma, "Synset Based multilingual dictionary: Insights, applications and challenges", *In Global Wordnet Conference,* Szeged, Hungary, 2008.

[26] Pushpak Bhattacharyya, "IndoWordnet", *Lexical Resources Engineering Conference 2010 (LREC 2010), Malta, May, 2010.*

About the Authors

Dr. Pushpak Bhattacharyya Dr. Pushpak Bhattacharyya is a Professor of Computer Science and Engineering at IIT Bombay. He received his B.Tech from IIT Kharagpur, M.Tech from IIT Kanpur and PhD from IIT Bombay. He has held visiting positions at MIT, Cambridge, USA, Stanford University, USA and University Joseph Fourier, Grenoble, France. Dr. Bhattacharyya's research interests include Natural Language Processing, Machine Translation and Machine Leaning. He has had more than 130 publications in top conferences and journals and has served as program chair, area chair, workshop chair and PC member of top fora like ACL, COLING, LREC, SIGIR, CIKM, NAACL, GWC and others. He has guided 7 PhDs and over 100 masters and undergraduate students in their thesis work. Dr. Bhattacharyya plays a leading role in India's large scale projects on Machine Translation, Cross Lingual Search, and Wordnet and Dictionary Development. Dr. Bhattacharyya received a number of prestigious awards including IBM Innovation Award, United Nations Research Grant, MIcrosoft Research Grant, IIT Bombay's Patwardhan Award for Technology Development and Ministry of IT and Digital India Foundation's Manthan Award. Recently he has been appointed Associate Editor of the prestigious journal, ACM Transactions on Asian Language Information Processing and also was chosen for Yahoo Faculty Award.