Semantic Role Labeling for Bengali Noun using 5Ws

Who, What, When, Where and Why

Amitava Das[1], Aniruddha Ghosh[2] and Sivaji Bandyopadhyay[3]
Computer Science and Engineering Department
Jadavpur University
Kolkata, India
amitava.santu@gmail.com [1] arghyaonline@gmail.com[2] sivaji_cse_ju@yahoo.com[3]

Abstract— **In this paper we present different methodologies to extract semantic role labels of Bengali nouns using 5W distilling. The 5W task seeks to extract the semantic information of nouns in a natural language sentence by distilling it into the answers to the 5W questions: Who, What, When, Where and Why. As Bengali is a resource constraint language, the building of annotated gold standard corpus and acquisition of linguistics tools for features extraction are described in this paper. The tag label wise reported precision values of the present system are: 79.56% (Who), 65.45% (What), 73.35% (When), 77.66% (Where) and 63.50% (Why).**

Keywords- Semantic Role Labelin; Bengali Noun, 5Ws

I. Introduction

In the last few years there has been an increased interest in shallow semantic parsing of natural languages as an important component in all kinds of Natural Language Processing (NLP) applications. Semantic Role Labeling (SRL) is a shallow semantic parsing technique that is now being widely used in question and answering (QA), information retrieval (IR) and extraction (IE), machine translation, paraphrasing, textual entailment, event tracking and so on.

The SRL task is to assign syntactic constituents (arguments) with semantic roles of predicates (most frequently verbs) at sentence level. A semantic role is the relationship that a syntactic constituent has with a predicate. Given a sentence, the task consists of analyzing the propositions expressed by some target verbs of the sentence. In particular, for each target verb all the constituents in the sentence that fill a semantic role of the verb have to be recognized. Typical semantic arguments include Agent, Patient, Instrument, etc. and also adjuncts such as Locative, Temporal, Manner, Cause, etc.

SRL has been extensively studied for English language but no such effort could be found in Indian languages and especially in Bengali. A linguistic annotation task for Hindi SRL is reported in [1]. The present work reports the development resources and methodologies to extract semantic role labels of Bengali nouns using 5W distilling.

The ideological study of semantic roles started age old ago since Panini's karaka theory that assigns generic semantic roles to words in a natural language sentence. Semantic roles are generally domain specific in nature such as FROM_DESTINATION,TO_DESTINATION,DEPARTURE_TIME etc. Verb-specific semantic roles have also been defined such as EATER and EATEN for the verb eat. The standard datasets that are used in various English SRL systems are: PropBank [2], [3] and [4]. These collections contain manually developed well-trusted gold reference annotations of both syntactic and predicate-argument structures.

PropBank defines semantic roles for each verb. The various semantic roles identified [5] are Agent, patient or theme etc. In addition to verb-specific roles, PropBank defines several more general roles that can apply to any verb [2].

FrameNet is annotated with verb frame semantics and supported by corpus evidence. The frame-to-frame relations defined in FrameNet are Inheritance, Perspective_on, Subframe, Precedes, Inchoative_of, Causative_of and Using. Frame development focuses on pa-raphrasability (or near paraphrasability) of words and multi-words.

VerbNet annotated with thematic roles refer to the underlying semantic relationship between a predicate and its arguments. The semantic tagset of VerbNet consists of tags as agent, patient, theme, experiencer, stimulus, instrument, location, source, goal, recipient, benefactive etc..

It is evident from the above discussions that no adequate semantic role set exists that can be defines across various domains. The idea of 5W semantic roles proposed in this paper aims to develop a generic semantic role set across domains and languages. The idea has been explored for Bengali language.

II. Historical Panini's Karaka Theory

The classical Sanskrit grammar *Astadhyayi* ('Eight Books'), created by the Indian grammarian Panini at a time variously estimated at 600 or 300 B.C. [6], includes a sophisticated theory of thematic structure that remains influential till today. Panini's Sanskrit grammar is a system of rules for converting semantic representations of sentences into phonetic representations [7]. This derivation proceeds through two intermediate stages: the level of *karaka* relations, which are comparable to the thematic role types described above; and the level of morphosyntax.

Grammar rules map each of the *karakas* to a basic semantic relation, and a basic morphosyntactic expression. More specialized variants of both types of rule are specified as well,

978-1-4244-6899-7/10/$26.00 ©2010 IEEE

with the basic relation and basic expression acting as defaults whenever the conditions for the variants are not met.

For example, the *karaka* called *apadana* (Source: Where/When) has as its basic semantic relation the fixed point from which something recedes. But with certain verbs *apadana* is instead used for special relations such as the source of fear, the object of hiding from, hindering, or learning from, and so on. The basic expression of *apadana karaka* is Ablative case. The basic semantic relation of the *karma karaka* (Theme: What) is that which is primarily desired; its basic expression is Accusative case. The *karana karaka* (Instrument: What) is associated with the basic semantic relation of the most effective means. While its basic expression is Instrumental case, some verbs are instead specified for the Genitive case to express the *karana* (such as '*break*', '*eat*', etc.) Other karakas include *sampradana* (Indirect Objec: What), *adhikarana* (Locative: Where), *karta* (Agent: Who), and *hetu* (Cause: Why).

Our aim and argumentation is for easy implementation of Panini's karaka theory at the crossroads of syntactic to semantic formalization of language aspects. However, on a closer look, several complications arise specially in Panini's recourse to semantics in many of the *vidhi* or *samajhna* rules. This seems to happen more in the *karaka prakarana* than in other components. An important effort [8] describes a syntactic annotation scheme for English based on Panini's concept of *karakas*.

The present work focuses on the semantic aspects of Panini's *karaka* theory using the simple and robust 5W distilling process and highlights the challenges in implementing these rules. Instead of using standard semantic role labels we use simple 5W concepts that can be easily mapped to Panini's *karaka* theory that robustly describes the syntactic and semantic synergy of any natural language.

III. Semantic Roles in Modern Generative Grammar

Fillmore's **Case Grammar** [9], and much subsequent work, revived Panini's proposals in a modern setting. A principle objective of **Case Grammar** was to identify semantic argument positions that may have different realizations in syntax. Fillmore hypothesized 'a set of universal, presumably innate, concepts which identify certain types of judgments human beings are capable of making about the events that are going on around them'. He posited the following preliminary list of cases, noting however that 'Additional cases will surely be needed' (and indeed Fillmore added more in later works [3]).

- Agent: The typically animate perceived instigator of the action. (**Who**)
- Instrument: Inanimate force or object causally involved in the action or state. (**What**)
- Dative: The animate being affected by the state or action. (***Who***).
- Factitive: The object or being resulting from the action or state. (**What**)
- Locative: The location or time-spatial orientation of the state or action. (**Where/When**).
- Objective: The semantically most neutral case conceivably the concept should be limited to things which are affected by the action or state. (**Why**)

IV. The Proposed Concept of 5Ws

In journalism, the Five Ws (Who, What, When, Where and Why) is a concept in news style, research and in police investigations that are regarded as basics in information gathering. The concept of 5Ws was first introduced by [10] in journalism.

- Who? Who was involved?
- What? What happened?
- When? When did it take place?
- Where? Where did it take place?
- Why? Why did it happen?

The 5W task seeks to summarize the information in a natural language sentence by distilling it into the answers to the 5W questions: Who, What, When, Where and Why. 5W concept is easy to understand in contrast with Panini's karaka theory or Fillmore's Case Grammar and even understandable by people have no linguistics expertization.

There are a small number of NLP applications where the ideas of 5Ws have been used successfully. In Machine Translation there is an evaluation methodology that uses the concept of 5Ws and addresses the cross-lingual 5W task: given a source language sentence it returns the 5Ws comprehensibly translated into the target language [11].

According to best of our knowledge the present work is the first attempt of Semantic Role Labeling in Bengali and in either aspect semantic information extraction using 5Ws distilling to map Panini's *karaka* theory. The 5Ws semantic role labeling task assigns domain independent generic semantic roles and should be considered as a supporting tag set to the kind of tags that are found in resources like Propbank, FrameNet or VerbNet. Such semantically tagged resources are very much necessary for several NLP applications in any language.

The 5Ws semantic role labeling task demands and addressing various NLP issues such as: predicate identification, argument extraction, attachment disambiguation, location and time expression recognition. To solve these issues the present system architecture relies on Machine Learning technique followed by a rule-based methodology.

One of the most important milestones in SRL literature is CoNLL-2005 Shared Task[1] on Semantic Role Labeling. All most all SRL research group participated in the shared task. System reports of those participated systems eminently prove that Maximum Entropy[2] (ME) based models work well in this problem domain as 8 among 19 systems used ME as the

[1] http://www.lsi.upc.es/~srlconll/st05/st05.html

[2] http://maxent.sourceforge.net/

solution architecture. The second best performing system [12] uses ME model uses only syntactic information without using any pre or post processing.

Table III presents the distribution pattern of 5Ws in overall corpus. It is very clear that 5Ws are not very regular jointly in the corpus as reported in Table III. Hence sequence labeling with 5Ws tags using ME will lead a label biased problem (as we reported in Section VI.A) and may not be an only acceptable solution for present problem definition as (Haghighi et al., 2005). The system, we proposed here follows a hybrid mechanism that statistically (ME based) assign 5W labels to each syntactic entity at sentence level and rule based post-processor helps to reduce many false hits by statistical system as well as identifies new 5W labels which increase the overall performance of the final system. The rule based post-processor works on the output of statistical system. The rules are being captured by acquired statistics on training set and linguistic analysis of standard Bengali grammar.

V. Resource Organization

Resource acquisition is one of the most challenging obstacles to work with resource constrained languages like Bengali. Bengali is the fifth popular language in the World, second in India and the national language in Bangladesh. Extensive NLP research activities in Bengali have started recently but resources like annotated corpus, various linguistic tools are still unavailable for Bengali. The manual annotation of the gold standard Bengali corpus is described in following section. The features to be found most effective are chosen experimentally. All the features that have been used to develop the present system are described in Feature Organization section.

A. *Gold Standard Data Acquisition*

1) *Corpus*

For the present task, the corpus from the ICON 2009 Dependency Parsing shared task[3] has been chosen. The data is manually annotated with part of speech (POS), chunk, morphological features and dependency tree relationships. Detailed reports about this corpus development in Bengali could be found in (Ghosh et al., 2009).

2) *Annotation*

The corpus statistics is presented in Table I. Sanchay[4], a well known linguistic annotation tool for Indian languages has been used for Bengali sentence level 5Ws manual annotation task. Two annotators (Mr. X and Mr. Y) participated in the present task. The annotated documents are saved in Shakti Standard Format[5] (SSF: XML format).

Annotators were asked to annotate 5Ws in Bengali sentences in terms of Bengali noun chunks. Instructions have been given to annotators to find out the principle finite verb in a sentence and successively extract 5W components by asking 5W questions to the principle verb. The annotators summarize the information in a natural language sentence by distilling it into the answers to the 5W questions: Who, What, When, Where and Why. An example of the 5Ws annotated document is presented in Figure 1.

TABLE I. Bengali News Corpus Statistics

Bengali Corpus Statistics			
	Train	**Dev**	**Test**
Total number of sentences in the corpus	980	150	150
Total number of wordforms in the corpus	9223	1762	1812
Total number of distinct wordforms in the corpus	6233	522	628

3) *Inter-annotator Agreement*

The agreement of annotations between two annotators has been evaluated. The agreements of tag values at each 5W level are listed in Table II.

TABLE II. Agreement of annotators at each 5W level

Tag	Annotators X and Y Agree percentage
Who	88.45%
What	64.66%
When	76.45%
Where	75.23%
Why	56.23%

It has been observed that in the present task the inter-annotator agreement is better for Who, When and Where level annotation rather than What and Why level though a small number of documents have been considered.

TABLE III. Sentence wise co-occurrence pattern of 5Ws

Tags	Percentage				
Who	What	When	Where	Why	Overall
	58.56%	73.34%	78.01%	28.33%	73.50%
What	Who	When	Where	Why	Overall
	58.56%	62.89%	70.63%	64.91%	64.23%
When	Who	What	Where	Why	Overall
	73.34%	62.89%	48.63%	23.66%	57.23%
Where	Who	What	When	Why	Overall
	78.0%	70.63%	48.63%	12.02%	68.65%
Why	Who	What	When	Where	Overall
	28.33%	64.91%	23.66%	12.02%	32.00%

[3] http://ltrc.iiit.ac.in/nlptools2009/

[4] http://sourceforge.net/projects/nlp-sanchay/

[5] http://web2py.iiit.ac.in/publications/default/download/techreport.pdf.c08a8d0a-50ed-4837-8ff0-93d099efbccb.pdf

```
<Sentence  id="1">
1        ((      NP      <fs  af='মাধবীলতা,n,,sg,,d,0,0'  head="মাধবীলতা",  Who>
1.1      মাধবীলতা          NN      <fs  af='মাধবীলতা,n,,sg,,d,0,0'  name="মাধবীলতা">
         ))
2         ((     VGNF    <fs  af='শো,v,,,2,,বে,be'  head="শোবে",  Why>
2.1      শোবে     VM      <fs  af='শো,v,,,2,,বে,be'  name="শোবে">
         ))
3         ((     VGF     <fs  af='বল্,v,,,3,,নে,ne'  head="বলে">
3.1      বলে      VM      <fs  af='বল্,v,,,3,,নে,ne'  name="বলে">
         ))
4         ((     NP      <fs  af='তখন,pn,,,,d,0,0'  head="তখন",  When>
4.1      তখন      PRP     <fs  af='তখন,pn,,,,d,0,0'  name="তখন">
         ))
5         ((     NP      <fs  af='হাত,n,,sg,,o,এর,era'  head="হাতের">
5.1      হাতের    NN      <fs  af='হাত,n,,sg,,o,এর,era'  name="হাতের">
         ))
6         ((     NP      <fs  af='ঘড়ি,unk,,,,,,'  head="ঘডি"  poslcat="NM",  What>
6.1      ঘড়ি      NN      <fs  af='ঘড়ি,unk,,,,,,'  name="ঘডি"  poslcat="NM">
         ))
7         ((     VGNF    <fs  af='খুল্,v,,,3,,নে,ne'  head="খুলে">
7.1      খুলে      VM      <fs  af='খুল্,v,,,3,,নে,ne'  name="খুলে">
         ))
8         ((     NP      <fs  af='টেবিল,n,,sg,,d,মে,me'  head="টেবিলে",  Where>
8.1      টেবিলে    NN      <fs  af='টেবিল,n,,sg,,d,মে,me'  name="টেবিলে">
         ))
9         ((     VGF     <fs  af='রাখ্,v,,,3,,ছিল,Cila'  head="রাখছিল">
9.1      রাখছিল   VM      <fs  af='রাখ্,v,,,3,,ছিল,Cila'  name="রাখছিল">
9.2      SYM     <fs  af='.,punc,,,,,,'  poslcat="NM">
         ))
</Sentence>
```

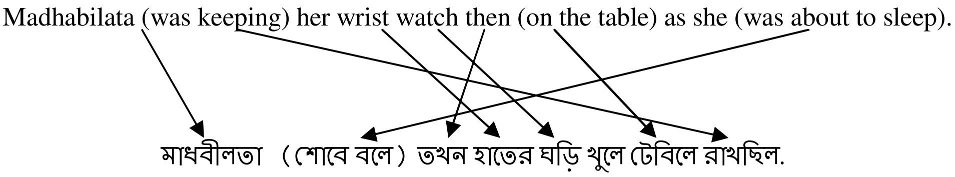
Madhabilata (was keeping) her wrist watch then (on the table) as she (was about to sleep).
মাধবীলতা (শোবে বলে) তখন হাতের ঘড়ি খুলে টেবিলে রাখছিল.

Figure 1. Example chunk level 5W annotated document

Further discussion with annotators reveals that the psychology of annotators is to grasp all 5Ws in every sentence, whereas in general all 5Ws are not present in every sentence. The observation is the most ambiguous tag to identify is "What". Let us take an example.

রাম/Who শ্যামের কানে/Where কানে/Where বললো।

Ram/Who whispers at Shyam's ear/Where.

In the preceding example Ram/রাম should be tagged as "Who" but it is ambiguous to find out the candidate for "What" tag. One annotator tagged Shyam's/শ্যামের as "What", but it is an animate object of the main verb whisper/বললো. Conceptually animate objects should be categorized as "Whom". In that case 5Ws should be listed as Who, What/Whom, When, Where and Why or 6Ws including "Whom". To disambiguate between "What" and "Whom" we created and corpus statistics for each 5W tag level listed in Table III.

It is shown in Table III that “What” occurrence is good as 64.23% in overall corpus. There are good number of cases where “What” is an animate object. But for the present task and for the sake of simplicity we only considered inanimate objects under “What” category.

Another important observation is 5W annotation task take very little time for annotation. Annotation is a vital tedious task for any new experiment, but 5W annotation task is easy to adopt for any new language.

B. *Feature Organization*

The set of features used in the present task have been categorized as Lexico-Syntactic, Morphological and Syntactic features. These are listed in the Table IV below and have been described in the subsequent subsections. The tool has been used here is Bengali Shallow Parser[6] developed under Indian Languages to Indian Languages machine Translation (IL-ILMT) project.

TABLE IV. FEATURES

Types	**Features**	
Lexico-Syntactic	POS	
Morphological	**Noun**	Root Word
		Gender
		Number
		Peson
		Case
	Verb	Root Word
		Modality
Syntactic	Head Noun	
	Chunk Label	
	Dependency Relation	

1) *Lexico-Syntactic Features*

a) *Part of Speech (POS)*

It has been shown by [13], [2] etc. that part of speech of any word in sentences is a vital clue to identify semantic role of that word.

2) *Morphological Features*

a) *Noun*

ROOT WORD

Root word is a good feature to identify word level semantic role especially for those types of 5Ws where dictionaries have been made like “When” and “Where”. To capture contextual dependency with particular post positional or conjunctive word root word of every chunk head has been kept in feature list.

GENDER

Gender information is essential to relate any chunk to the principle verb modality. In the case of “What”/”Whom” ambiguities gender information help significantly. For inanimate objects it will be null and for animates it has definitely a value. Bengali is not a gender sensitive language hence this feature is not such significant linguistically rather number and person features. But the statistical co-occurrence of gender information with the number and person information is significant.

NUMBER

Number information help to identify specially for “Who”/”What” ambiguities. As we reported in inter-annotator agreement section “Who” has been identified first by matching modality information of principle verb with corresponding number information of noun chunks.

PERSON

Person information is as important as number information. It helps to relate any head of noun chunks to principle verb in any sentence.

CASE

Case markers are generally described as karaka relations of any noun chunks with main verb. It has been described that semantically karaka is the ancestor of all semantic role interpretations. Case markers are categorized as Nominative, Accusative, Genitive and Locative. Case markers are very helpful for almost in every 5W semantic role identification task.

3) *Syntactic Features*

a) *Head Noun*

The present SRL system identifies chunk level semantic roles. Therefore morphological features of chunk head is only important rather other chunk members.

b) *Chunk Label*

Present SRL system identifies noun chunk level semantic roles. Hence chunk level information is effectively used as a feature in supervised classifier and successively in rule-based post processor.

c) *Dependency Parser*

It has been profoundly established that dependency phrase-structures are most crucial to understand semantic contribution of every syntactic nods in a sentence [13], [2]. A statistical dependency parser has been used for Bengali as described in [14].

VI. SEMANTIC ROLES IDENTIFICATION

A. *Using MEMM*

MEMM treats 5Ws semantic role labeling task as a sequence tagging task. MEMM views the problem as a pattern-matching task, acquiring symbolic patterns that rely on the syntax and lexical semantics and morphological features of a

[6] http://ltrc.iiit.ac.in/analyzer/bengali/

phrase head. With the all selected features properties in mind and supported by series of experimentation, we finalize the final features(described in Table IV) for each chunk level in an input sentence. For pedagogical reasons, we may describe some of the features as being multi-valued (e.g. root word) or categorical (e.g. POS category) features. In practice, however, all features are binary for the MEMM model. In order to identify features we started with Part Of Speech (POS) categories and continued the exploration with the other features like chunk, Dependency relation and morphological features.

TABLE V. PERFORMANCE OF 5WS SRL BY MEMM

Tag	Precision	Recall	F-measure	Avg F-Measure
Who	76.23%	64.33%	69.77%	62.22%
What	61.23%	51.34%	55.85%	
When	69.23%	58.56%	63.44%	
Where	70.01%	60.00%	64.61%	
Why	61.45%	53.87%	57.41%	

The feature extraction pattern for any Machine Learning task is crucial since proper identification of the entire features directly affect the performance of the system. 5Ws Semantic role labeling is difficult in many ways. A sentence does not always contain all 5Ws. Although Bengali is defined as a verb final language but there is no certain order in occurrence among these 5Ws in a sentence. The performance of 5W SRL task by MEMM is reported in Table V.

It is noticeable that the performance of the MEMM-based model differs tag-wise. While precision values for “Who”, “When” and “Where” is good but recall yielded i.e. system failed to identify in various cases. In “What” cases system identified most of the cases as recall is high but also make so many false hits as precision is low. For tag label “Why” precision and recall values both are low as reported.

For such heterogeneous problem nature we propose a hybrid system as rule-based post processor followed by Machine Learning. The rule-based post processor can identify those cases missed by ML method and can reduce false hits generated by statistical system. These rules are formed by heuristic on gold standard data and standard Bengali grammar.

B. Rule-Based Post-Processing

1) Who? Who was involved?

As described earlier system failed to identify “Who” in many cases. As an example:

নিমন্ত্রিত না হলেও তোমার/Who যাওয়া উচিত ছিল সেখানে।

Though you are not invited but you/Who should go there.

System fails in this type of cases, because the targeted chunk head is a pronoun and it is situated at almost in the middle of the sentence whereas “Who” generally situated at initial positions in a sentence as Bengali Verb final and Subject initial language. Moreover system made some false hits too. As an example:

দরজাটা বন্ধ করো।

Close the door.

In the previous case system mark দরজাটা/door as a “Who” whereas the “Who” is “you” (2nd person singular number), silent here. This an perfect example of label-biased problem. System is quite biased towards those chunks at initial position of sentences.

We developed rules using case marker, Gender-Number-Person (GNP), morphological subject and modality features to disambiguate these types of phenomena. These rules help to stop false hits by identifying no 2nd person phrase was there in the type of second example sentences and empower to identify proper phrases by locating proper verb modality matching with the right chunk. These rules increase system overall performance value reported in Section VII.

2) What? What happened?

As described in earlier sections “What” could be also described as “Whom” where object is animate. To avoid further ambiguities we categorize both animate and inanimate objects as “What” for the present task. The corpus distribution of “What” and “Whom” is almost 50-50% ration as we noticed. In the next examples in the first sentence বাঁশি/ Flute is semantically “What” whereas in the next sentence তাকে/him is representing semantic “Whom”.

শ্যামের বাঁশি।

Flute of Shyam.

এটা তাকে দিও।

Gave this to him.

We make use of only positional information for “What” or object identification. There is less syntactic, orthographic and morphological difference between “Who” and “What”. For that reason a reduction methodology has been used as “Who” has been detected by system first and “What” has been tagged among rest of the noun chunks with positional factor in the sentence.

Significant increment in result could be noticed in Section VII.

3) When? When did it take place?

As addressed in Introduction section time expression identification has a different aspect in NLP applications. People generally studied time expression to track event or any other kind of IR task. We incorporate this for SRL. Time expressions could be categorized in two types as General and Relative as listed in Table VI.

In order to apply rule-based post-processor we developed a manually augmented list with pre defined categories as described in Table VI. Still there is many difficulties to identify special cases of relative time expressions. As an example:

চাঁদ উঠলে আমরা রওনা হবো।

When moon rise we will start our journey.

TABLE VI. CATEGORIES OF TIME EXPRESSIONS

	Bengali		English Gloss
General	সকাল/সন্ধ্যে/রাত/ভোর…		Morning/evening/ night/dawn…
	_টার সময়/সময়/ঘটিকায়/মিনিট/সেকেন্…		O clock/time/hour/ minute/second…
	সোমবার/মঙ্গলবার/রবিবার…		Monday/Tuesday /Sunday…
	বৈশাখ/জৈষ্ঠ/…		Bengali months…
	জানুয়ারী/ফেব্রুয়ারী		January/February …
	দিন/মাস/বছর…		Day/month/year …
	কাল/ক্ষন/পল…		Long time/moment…
Relative	আগে/পরে…		Before/After…
	সামনে/পেছনে…		Upcoming/
	Special Cases	উঠলে/থামলে…	When rise/When stop…

In the previous example the relative time expression is উঠলে/when rise is tagged as infinite verb (for Bengali tag level is VGNF). But the scope of the present system is only nouns hence this types of cases arre not handled currently. Statitics reveals that these special type of cases approximately are only 1.8-2% in overall corpus.

As like "Who" these manually augmented list followed by some hand crafted rules increase the rage of identification of "When" in Bengali sentences. Performance increment in recall value is reported in Section VII.

4) Where? Where did it take place?

Identification of "Where" simply refers to the task of identification locative marker in NLP. As "When", we categorized "Where" as general and relative as listed in Table VII.

Rules have been written using a manually edited list as described in Table VII. Morphological locative case marker feature have been successfully used in identification of locative marker. There is a ambiguity among "Who", "When" and "Where" tag as they orthographically generates same type of surface form (using common suffixes as: ে, ের etc). There is less differences we noticed among their syntactic dependency structure throughout corpus.

দেশে কাজ নেই বাবু।

There is unemployment in country side.

TABLE VII. CATEGORIES OF LOCATIVE EXPRESSIONS

	Bengali	English Gloss
General	মাঠে/ঘাটে/রাস্তায়	Morning/evening/night/dawn …
Relative	আগে/পরে…	Before/After …
	সামনে/পেছনে …	Front/Behind

For same kind of orthographic structure and morphological or syntactic affinity ML based model assign "Who" tag to দেশে/country side. Positional information is not helpful in previous example as দেশে situated in initial position of the sentence. Hence rules have been formulated using only morphological locative marker.

A different type of problem we found where verb plays "Where" semantic role. As an example:

লোকে যেখানে কাজ করে সেখানে।

Where people works there.

Here যেখানে কাজ করে/Where people works should be tagged as "Where". But this is a verb chunk and present scope of our work is only noun. Corpus statistics reveals that this type of syntactic formation is approximately 0.8-1.0% only.

Significant change in performance reported in Section VII.

5) Why? Why did it happen?

TABLE VIII. CATEGORIES OF CAUSATIVE EXPRESSIONS

	Bengali	English Gloss
General	জন্য/কারনে/হেতু…	Hence/Reason/ Reason
Relative	যদি_তবে	If_else
	যদিও_তবুও	If_else

The particular role assignment for "Why" is the most challenging task as it separately known as argument identification. As reported in previous sections inter-annotator agreement and overall distribution regularity is very low. For irregular and small occurrence of "Why" leads poor result in ML-based technique. Inter-annotator agreement shows that even human annotators are also disagree about the "Why" tag. To resolve this problem we need a relatively large corpus to learn fruitful feature similarities among argument structures.

A manually generated list of causative postpositional words and pair wise conjuncts as reported in Table VIII has been prepared to identify argument phrases in sentences.

Small incremental changes could be no-ticed in precision value of “Why” identification but no significant recall has been noticed as reported in Section VII.

VII. Experimental Result

The performance result of ML technique has been reported in Table V. After using rule-based postprocessor the system performance increases as listed in the following Table IX.

TABLE IX. Performance of 5Ws SRL by MEMM+Rule-Based-Post Processing

Tag	Precision	Recall	F-measure	Avg F-Measure
Who	79.56%	72.62%	75.93%	68.10%
What	65.45%	59.64%	62.41%	
When	73.35%	65.96%	69.45%	
Where	77.66%	69.66%	73.44%	
Why	63.50%	55.56%	59.26%	

VIII. Conclusion and Future Works

In this paper we described a novel approach to assign semantic roles of Bengali nouns by 5W distilling. According to best of our knowledge this is the first attempt of information extraction using 5Ws distilling specifically in Semantic Role Labeling for Bengali nouns.

To avoid the debate of using of very specific type tagset or in general type tagset 5Ws could give acceptable solution architecture. Generally in depth semantic role labeler is not hard requirement for all type of NLP applications like: IR, IE, Textual Entailment, Multi-Document Summarization etc. Usage and the develop-ment of an in depth SRL is nothing but spill over with information for those applications moreover development of such heavy weight SRL for a resource constraint lan-guage like Bengali required hectic manual annotation and other standard well trusted linguistic resources like WordNet, VerbNet, FrameNet and PropBank etc.

The proposed 5W concept is not a global solution for all type of semantically interpreted NLP applications. Rather it has been proposing that the 5W concept is very simple but effective for more or less all kind of modern NLP applications. Development of heavy weight SRL will be still in horizon to reach and required to discover in depth semantic interpretation for any new resource constraint Indian languages like Bengali.

We are presently studied SRL techniques for other part-of-speech categories. We are planning to apply present SRL to some NLP application area like opinion mining, event tracking or textual entailment to discover its real feasibility in real life problem domain.

References

[1] Sapna Sharma, Karthink Gali and Soma Paul. Au-tomatic annotation of semantic relations on Noun Genitive constructions in Hindi Language. In ICON-2009: 7th INTERNATIONAL CON-FERENCE ON NATURAL LANGUAGE PROCESSING. Hyderabad, India.

[2] Martha Palmer, Dan Gildea, Paul Kingsbury, The Proposition Bank: A Corpus Annotated with Semantic Roles, Computational Linguistics Journal, 31:1, 2005.

[3] Charles J. Fillmore, Christopher R. Johnson, and Miriam R. L. Petruck. 2003. Background to FrameNet. International Journal of Lexicography, 16:235–250.

[4] Karin Kipper, Anna Korhonen, Neville Ryant, and Martha Palmer. Extending VerbNet with Novel Verb Classes. Fifth International Conference on Language Resources and Evaluation (LREC 2006). Genoa, Italy. June, 2006.

[5] Dowty, David R. 1991. Thematic proto-roles and argument selection. Language, 67(3):547–619.

[6] Robins, R. H. (1979). A Short History of Linguistics (2nd Edition). London: Longman.

[7] Kiparsky, Paul and J. F. Staal (1969). ‘Syntactic and semantic relations in Panini.’ Foundations of Language 5, 83-117.

[8] Ashwini Vaidya, Samar Husain, Prashanth Mannem, and Dipti Misra Sharma. 2009. A Karaka Based Annotation Scheme for English. In Proceedings of the 10th International Conference on Computational Linguistics and Intelligent Text Processing. Lecture Notes In Computer Science; Vol. 5449. Pages: 41 – 52. ISBN:978-3-642-00381-3.

[9] Fillmore, Charles (1968). ‘The Case for Case.’ In Emmon Bach and Robert T. Harms (eds.) Universals in Linguistic Theory. New York: Holt, Rinehart, and Winston. 1-88.

[10] Philip F. Griffin, "The Correlation of English and Journalism" The English Journal 38:4 (April 1949), pp. 192 at JSTOR.

[11] Kristen Parton, Kathleen R. McKeown, Bob Coyne, Mona T. Diab, Ralph Grishman, Dilek Hakka-ni-Tür, Mary Harper, Heng Ji, Wei Yun Ma, Adam Meyers, Sara Stolbach, Ang Sun, Gok-han Tur, Wei Xu and Sibel Yaman. Who, What, When, Where, Why? Comparing Multiple Ap-proaches to the Cross-Lingual 5W Task. In the Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP, pages 423–431, Suntec, Singapore, 2-7 August 2009.

[12] Aria Haghighi, Kristina Toutanova and Christopher D. Manning. A Joint Model for Semantic Role Labeling. In CoNLL-2005 Shared Task.

[13] Daniel Gildea and Daniel Jurafsky, Automatic Labeling of Semantic Roles, In Association for Computational Linguistics, 2002.

[14] A. Ghosh, A. Das, P. Bhaskar, S. Bandyopadhyay. Dependency Parser for Bengali: the JU System at ICON 2009, In NLP Tool Contest ICON 2009, December 14th-17th, 2009, Hyderabad.