Information Extraction from Hindi Texts

Kamlesh Dutta*, Saroj Kaushik, Nupur Prakash*****

*National Institute of Technology, Hamirpur (HP) INDIA 177005
kd@recham.ernet.in
** Indian Institute of Technology, New Delhi INDIA
saroj@cse.iitd.ernet.in
*** IndiraGandhi Institute of Technology, New Delhi, INDIA
nupurprakash@rediffmail.com

Abstract

The paper presents an information extraction system that takes input from Hindi texts and improves the information content retrieved by using anaphor/pronoun resolution mechanism. The information extraction system developed consists of three major modules: The language Parser, Resolution System and Information Extractor. The language parser used is HPSG (Head-Driven Phrase Structure Grammar) based that provides both syntactic and semantic information to the anaphor resolution system. HPSG was chosen because it provides a set of constraint on the co-referential structures in the language, which bounds the search for an antecedent to a more precise location in the discourse. The semantic information included in its parsing may be helpful for removing ambiguity in anaphor/pronoun resolution. The anaphor resolution system uses few heuristic rules to resolve intrasentential references while centering theory is used for intersentential resolution

Introduction

Information Extraction (IE) has been used to develop specific software's for automatic summary generation, email processing to answer database queries in Natural Language [4] and simple question answer systems. In the Indian context, for the successful deployment of Information technology, there is a need to develop tools for Indian language. It can prove a major pitfall in the IT vision of our country if we let ourselves lag behind in a promising field of IE. This is more so because of very low level of English literacy in our country. For Hindi Language (Dave and Bhattacharya 2001) have used Universal Network Language for knowledge extraction from Hindi text which preserves the predicate till the end. The result of the analysis is the semantic net like structure.

The Information Extraction system developed in the present work for Hindi has following main modules:

- Language Parser
- Anaphor/Pronoun Resolution System
- Information extractor

Hindi Language Parser

The Language Parser reads the input Hindi text provided and checks it for error. The parser is based on Head Phrase Structured Grammar (Pollard and Sag, 1994) . The output of HPSG parser will contain parts of speech tagging, the number, gender and case specification and semantic information for phrases. This information is arranged as a Case Frame for every sentence, which contains the syntactic and semantic information of the sentence.

HPSG structure used for representation is ideally suited for anaphor resolution because of its constraint-based nature and since it divides the sentence into a hierarchical arrangement of Head and phrase as shown by (Pollard and Sag, 1994),(Pollard 1996). This hierarchy is useful to resolve anaphors and pronouns whose antecedents are bounded according to the phrase of occurrence of the anaphors.

Anaphora/Pronoun Resolution

Using the semantic and syntactic information from the Case Frame Structure, the Anaphor resolution system tries to link the pronouns and anaphors to their referents. Each noun/pronoun object in the case frame structure has an INDEX field and a REFERENT_INDEX field. In the input the NPs/anaphors are assigned a unique index number by the parser. During the passes of resolution system, the REFERENT_INDEX field of referring objects is set to the index of referent. Some anaphors/pronouns may remain unresolved, which are assigned a REFERENT_INDEX value of zero.

The anaphor resolution system consists of two major jobs:
Anaphor Resolution: According to the Binding theory for HPSG grammar an anaphor must be bound in its Governing Category (Grosz et al 1995). Heuristic rules have been developed to identify the referent object. Simple Heuristics like Gender and Number may be used to resolve and verify the link if there are more than one prospective object (Pirkola and Jarvelin1994) and also shown by (Sobha, and Patnaik 2000) for anaphor resolution in Hindi Language. Centering theory provides the list of most probable NPs which may be the referents of a pronoun in a given sentence. At any occurrence of a pronoun, the probable list is considered in decreasing order of importance and various heuristics are applied to resolve the pronoun to one of the NP in probable list

Information Extractor

The output of the Anaphora Resolution System is a Mapping Table that links the anaphors/Pronouns to their referents. This module then attempts to infer the meaning of the sentences. It finds the logical relations that hold between the object, the events that occur and the object taking part in the events. The information can be represented as prolog predicates, which can then be used for various purposes like Machine Translation and Reasoning.

e.g. consider the following discourse

Shyam is a student. He goes to college. Name of his college is NIT Hamirpur.

This set of sentences would yield us following information:

Student(Shyam).
Goes(Shyam, College).
College(NIT Hamirpur).
Belongs_to(NIT Hamirpur, Shyam).

HPSG is a constraint-based grammar. It mainly defines the syntactic and semantic rules to be followed by any grammatical construct. HPSG structures may be of following types

{ sign, word, phrase, category, Head (=part of speech), list, set, content, case, index, verb_form etc}.

Consider the verb 'rakkhi' in the following sentence

kitaab mej par rakkhi hai.

The HPSG structure for this verb is as shown in fig. 1:

```
|           |HEAD      | Verb     |                                   |  |
|           |          |  Non_Aux |                                   |  |
| CATEGORY  |          |          |                                   |  |
|           |                                                          |  |
|           | VALANCE  |SUBJ  :<NP[nom, inanimate]::[1]>             |  |
|           |          |COMP :<NP[locative,place] ::[2] ,             |  |
|           |          |               PP[locative]      ::[3]>       |  |
|                                                                       |  |
|           | Rakkhi            |                                        |
|           |OBJECT       [1]   |                                        |
| CONTENT   | LOCATION   [2]    |                                        |
|           |                   |                                        |
```

Fig1: HPSG Representation of Hindi Verb

From this example we see that in order to develop HPSG specification for Hindi we must have following information:

1. A lexicon containing the atomic objects and their properties. i.e. we must store in our knowledge base the proper and common nouns, different type of pronouns, prepositions, conjunctions. The basic characteristics in various contexts should also be stored eg. Whether a noun is of type animate, inanimate, place, time, event etc..
2. All verbs must be stored in the lexicon with information about the number and type of arguments required by them.
3. All semantic actions and events must be listed with the objects (and their types), which take part in the event with various roles.

Parsing would then consist of

1. Reading the words in sentences and identify the root word.
2. According to the form of root word used in the sentence, construct the HPSG word structure for this instance with the help of the structure template stored in knowledgebase.

After all individual words have been parsed, identify the arguments for the head of various constructs according to the constraints and index them accordingly. The head of phrase will automatically have the words as arguments while the head of sentence will have various phrases as its arguments thus completing the hierarchical structure.

Hindi has large number of pronouns. They cannot be clearly identified just on the basis of the word. Same word can be used as a pronoun in one place while it can be a demonstrative adjective at other place. This anomaly is assumed to have been sorted by the parser before anaphora / pronoun resolution.

Eg. Vah pustak meri hai. (vah is demonstrative adjective)

Vah mej par rakkhi hai. (vah is pronoun)

While certain pronouns are pure anaphors according to Binding theory ie they have their antecedent within their domain, others may be bounded outside their own domain. Eg. words like 'Apna' and 'swayam' are anaphors while 'vah','jisne','usne' etc are pronouns.

Ram ney **use** bulaya. (pronoun use used for some second person)

Ram ney **apne** bhai ko bulaya. (anaphor apne used for Ram)

Hindi sentences can be represented according to HPSG requirements. We have chosen to represent the information of the sentences in a case frame structure which conveys sufficient information for anaphor resolution and information extraction systems and is also simpler. It has flatter organisation of participating VP, NP, adjectives, adverbs and phrases/subsentences as compared to deep-rooted hierarchical organisation of HPSG.

The case frame structure of a sentence has following fields:

TOKEN: The text of the sentence for purpose of reproducing the sentence if required.

ID : A unique identifier of a sentence assigned by the parser. Each sentence and subsentence related to them have a unique ID.

VERB:

The verb phrase further contains fields to specify the type and property of the main verb in the sentence. In some cases where there are auxiliary verbs also present , they are clubbed with the main verb because the sentence can always be rearranged in such a manner as to convey the same meaning, preserve the syntactic structure and have a single combined verb.

e.g. Ram khana kha_kar mandir gaya.
Ram mandir kahana kha_kar gaya.

The verb phrase further has following fields to provide more information:

- TOKEN : Specifies the exact word as its occurance in the sentence
- ROOT: tells the basic form of the verb.
- TYPE of verb i.e. its transitivity etc.
- TENSE describes the tense of the event/action in the sentence.
- List of adverbs : it is a list of all adverbs of the main verb

These are the features currently incorporated because they are of help in the anaphor resolution system. But it is very flexible and new information fields can be included without affecting the present logic for anaphor resolution. In the future implementation the verb phrase may contain more semantic knowledge as specified by HPSG.

We can have a knowledge base storing the number and type (animate, inanimate, place, event etc.) of noun phrases which are needed as argument to every verb. A di-transitive verb for example will have two noun phrases. The parser must recognize the verb in the sentence, find its transitivity and the corresponding noun phrases. This information must then be entered into the case frame structure for further analysis.

Such information will specially be required by the Information extraction system which must know the transitivity and the arguments of a verb.

A typical verb phrase representation in case frame is as follow:

VERB: [TYPE:[di-transitive], TOKEN:[bulaya],
TENSE:[past] SEQUENCE:[3],]

SUBJECT:

Subject is a noun phrase, which is the cause, or the initiator of the event described in the sentence. It is the main focus Entity of the sentence. Every sentence must have a subject. If there is no explicit subject that can be recognized by the parser then the sentence has a Zero anaphor. The Parser must in that case insert a dummy NP as subject.

e.g. Ram ney kitab uthai aur [_dummy_] school chala gaya.

A NP 'sunderta' in the following sentence is represented in Case frame as follows:

Tajmahal ki sunderta adbhut hai.

SUBJECT: [TYPE:[abstract], TOKEN:[sunderta], ROOT:[sunder],NUMBER:[singular], GENDER:[female], CASE:[nominative], EXTENSION:[], SEQUENCE:[5], INDEX : [3] ADJ:[TYPE:[qualitative], TOKEN:[adbhut], NUMBER:[singular], GENDER:[male]]]

Every noun phrase representation including that of the subject contains following information :

- TOKEN : the exact occurance of the word in the sentence.
- TYPE : the type of noun (proper, collective, abstract..) if it is a noun or its value is set to indicate that it is a pronoun.
- ROOT : the basic form of the noun or pronoun without any number, gender, or case induced change.
- NUMBER: it can have value 'singular' or 'plural' as identified by the parser.
- GENDER: It can have value 'male' or 'female'. In case of pronouns or nouns whose gender cannot be inferred by the word or the verb of sentence, it is set to null.
- CASE : Defines the relation of the noun phrase with the main verb of the sentence. It can have one of eight values given in Table 8.
- EXTENSION : Extension of a noun includes those words which are used only for further description of the noun and are not covered by any other part of speech.
- SEQUENCE : Though Hindi is largely a word order free language but sometimes the placement of pronouns and noun is significant for their relation. Eg. if a NP occurs in Genetive case then it must be followed by the NP it is linked to.

Eg.

1 2 3 4 5
Ram ney Shyam ko [**uski** $_{NP}$] [**purani pustak** $_{NP}$] di.

Sequence value shows the position of occurrence of NP and VPs in the sentence. Only NP, pronouns, and main verb have a sequence value. The sequence value of the connector (conjunction, and few words like 'ki', 'kyonki', 'jabki' etc.) specifies the position of the beginning of a subsentence.

- INDEX : Index value of a NP is a unique identifier assigned by the parser. It is used to refer to the NP in the argument list of verb. The pronouns and anaphors are resolved by mapping their INDEX values with that of their antecedent.

Further improvements in this representation of NP are possible to include more semantic and world knowledge. For example the type of object ie whether it is animate, inanimate, event, property, place etc. Also the same NP may have different meaning in different sentences depending on the context. eg. 'kal' refers to 'Tomorrow' or 'Yesterday' when referring to time while it also means 'Machine Parts' when used in context of machinery. So a knowledgebase must store various meanings of the NP in different contexts. The parser can identify the context and meaning by analyzing the Verb and other constraint as specified in HPSG structure of the VP in sentence.

Subsentences:

Subsentences in compound or mixed sentences are represented under the CONNECTOR construct. A subsentence is modeled exactly like a simple sentence except that it is embedded at a lower hierarchy than its main sentence. The subsentence has its own ID different to the parent sentence's identification number. There may be more than one subsentences in a sentence which will be represented as sequence of CONNECTOR constructs.

An example of a Compound Sentence representation using case frame structure is as follow:

[TOKEN:[Tajmahal ek sunder bhavan hai,] ID:[1] VERB: [TYPE:[?], TOKEN:[hai], TENSE:[present] SEQUENCE:[3],], SUBJECT: [TYPE:[proper], TOKEN:[Tajmahal],ROOT:[Tajmahal], NUMBER:[singular], GENDER:[male], CASE:[nominative], EXTENSION:[], SEQUENCE:[1], INDEX[1]],OBJECT:[TYPE:[common],TOKEN: [bhavan],ROOT:[bhavan],NUMBER:[singular], GENDER: [?], CASE:[objective] EXTENSION:[], SEQUENCE:[2], INDEX : [2], ADJ:[TYPE:[number], TOKEN:[ek], NUMBER:[singular], GENDER:[male]], ADJ: [TYPE:[qualitative], TOKEN:[sunder], NUMBER:[singular], GENDER:[male]]] CONNECTOR: [TOKEN:[jiski sunderta adbhut hai.] ID:[2], VERB: [TYPE:[?], TOKEN:[hai], TENSE:[present] SEQUENCE:[6],],SUBJECT:[TYPE:[abstract],TOKEN:

[sunderta], ROOT:[sunder], NUMBER:[singular],
GENDER:[female], CASE:[nominative],
EXTENSION:[], SEQUENCE:[5],
INDEX:[3]ADJ:[TYPE:[qualitative],
TOKEN:[adbhut], NUMBER:[singular],
GENDER:[male]]], OBJECT: TYPE:[pronoun],
TOKEN:[jiski], ROOT:[jo], NUMBER:[singular],
GENDER: [male],CASE:[genative], EXTENSION:[],
SEQUENCE:[4], INDEX : [4]]]]

Representation of Sentences

The case frame structure used in our representation of Hindi sentences can be directly modeled into corresponding C language structures. The structures used to represent Noun Phrases, Verb Phrases, Adjectives and
In Hindi a verb may have up to eight different types of Objects related to it. i.e. the transitivity of the verb can be up to eight. At least one of the objects must be Subject.
Eg.
Ram ney **Ravan** ko **marney** ke liye **rath** sey utarkar **teer** sey **Ravan** ke **sir** par mara.
In the case frame structure the Subject is represented as a separate NP and all other argument object of the verb are listed as sequence of NPs. This arrangement suggests the use of an Array of NP objects along with an integer to specify the actual number of objects.
So a Sentence Object (structure) will contain a VP object, a NP object as subject, an Array of NPs for other Objects, and an integer to specify the actual number of Objects present.
The sub-sentences or phrases are implemented as a link to another Sentence Object. So each Sentence Object also contains a pointer to its sub-sentence object.
Such a data structure mirrors the Case frame structure and preserves the hierarchical arrangement of adjectives with their NP, Object and Subject with their VP and subsentences with their parent sentence.

Discourse Specification:

A discourse consists of number of related sentences. According to Centering theory(Grosz et al 1995), a discourse, utterances are connected to (related to) each other semantically by the center concept (or centers). In our implementation of discourse, the sentences are represented by a list of sentences. They are implemented by an array of pointers to Sentence Objects in sequence of their occurrence. The centering information is stored in a Data structure that contains a list of recent objects (NPs) and the weights assigned to them. The weights represent the relative importance of the NPs in the discourse. The NP is added into the list as and when they occur in the discourse. If the NP already exists then the weight of that NP in the sentence is added to the existing weight of NP in the List. After processing every sentence, the existing weight of the objects is decreased by a factor. We have chosen the factor to be 40%. The reason for this is not theoretical rather practical because it has worked well with most commonly occurring discourse in Hindi (Prasad and Strube 2000). A NP is removed from the list if its weight falls below a specified minimum value.

Evaluation

The anaphor approach used is tested over 10 short stories and following accuracy was observed:

Correct resolution: 63%
Correct third person pronoun resolution: 69.2%
Correct Definite pronoun resolution: 0%
Correct Zero Pronoun resolution: 100%
Correct Inter-sentential Pronoun Resolution: 54.5%
Correct Intra sentential Pronoun Resolution: 87.5%

The results suggested that the use of pronoun resolution improves the information content to be extracted which otherwise be ignored. However the algorithm has to be tested on different categories of texts also.

Conclusion

The Information extraction system for Hindi texts developed here uses heuristic approach to resolve the anaphors and pronoun. The rules used are applicable for most occurrences of pronouns in natural Hindi text. This is especially useful in descriptive texts, which have fewer occurrences of first and second person pronouns, which are not covered by the heuristics suggested. HPSG will add more semantic information and semantic constraints into the representation making the resolution more accurate.
The next step after Information extraction in Hindi texts is to extend it for web related text. A major problem in this regard is the absence of any standard encoding for Hindi alphabets. Various websites use proprietary font families to display same text. (eg Amarujala.com uses 'au' font family while dainikjagran.com uses 'jagran' family of font).

References

Dave Shachi and Bhatacharyya P.(2001) Knowledge Extraction from Hindi. Journal of Institution of Electronic and Telecommunication Engineers, vol. 18, no. 4, July, 2001.

Grosz, B.J, Joshi, A.K., and Weinstein, S. (1995). Centering: A framework for modelling the local coherence of discourse. Computational Linguistics, 21; (pp. 203-225)

Pirkola, A. and Jarvelin,K. (1994) The Effect of Anaphora and Ellipsis Resolution on Proximity Searching in a Text Databases. University of Tampere, Department of Information Studies, RN-1994-1, 25 p.

Pollard, C.(1996). HPSG: An Overview and some work in Progress. Pacific Asia conference on Languages, Information and Computation. Kyung Hee University, South Korea

Pollard,C and Sag,I.E.(1994) Head-Driven Phrase Structure Grammar. University of Chicago Press and Stanford: CSLI Publications

Prasad R. and Strube M.(2000) Discourse Salience and Pronoun Resolution in Hindi. In Penn working Papers in Linguistics, Vol 6.3(pp. 189-208)

Sobha, L. and Patnaik, B.N.(2000) Vashisht: An anaphora resolution System for Malayalam and Hindi. In International Conference ACIDCA'2000, Monastir, Tunisia