Volume 3, Issue 3, March 2013 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Rule Based Semantic Parsing Approach for Kashmiri Language

Manzoor Ahmad
Department of Computer Sciences,
University of Kashmir
Srinagar, Kashmir -190006, India

Romana riyaz
Department of Computer Sciences,
University of Kashmir
Srinagar, Kashmir -190006, India

Abstract-This paper describes a rule based semantic parsing approach for Kashmiri language(multilingual language) .A rule based semantic parser relies on the WORDNET and FRAMENET available for the language. Semantic parsing involves mapping of natural language (NL) sentence into a complete formal meaning representation (MR) in a meaning representation language. Semantic parsing increases the level of understanding of natural language processing than the syntactic parsing which is primarily involved in solving syntactic ambiguities by recursively applying sequences of context free rewriting rules.

Index terms- FrameNet, Semantic role assignment, Word Net, Lexicon, ontology, Syntactic.

I. Introduction

Semantic parsing is the process of mapping a natural language sentence into a formal representation of its meaning. A shallow form of Semantic representation is a case-role analysis (semantic role labeling).A deeper semantic analysis provides a representation of the sentence in predicate logic or other formal language which supports automated reasoning. Traditionally semantic parsers were constructed manually but this is too costly and brittle. Now a day's a number of machine learning approaches have been proposed [10].

The semantic parsers work in the same spirit as syntactic parsers which parse a natural language input into a parse tree and decompose a sentence syntactically. While as a semantic parser analysis the sentence in terms of its meaning. This task requires access to representations that link the linguistic elements involved in the input to the non-linguistic knowledge of the world. Sentence meaning is composed of entities and interactions between the entities. Roles are associated with semantic entities which can be modified by modifiers. The basic steps for any natural language understanding application is the syntactic analysis. There are some of the complications that inevitably arise in the process of designing meaning representations like verifiability, unambiguous representations, canonical forms etc. The semantic parser attempts to solve this problem and produces a syntax independent representation of sentence meaning and freeing it from the rigid interpretations produced by a syntactic analyzer. In this paper we make use of WordNet and FrameNet available as a starting point for rule based semantic parser.

II. Semantic Structure

The general principle of the compositionality of semantics is that the meaning of the whole is the function of the meaning of the parts. Meaningful components correspond to syntactic constituents.
Syntactic rule: *S -> C*1,*C2; ...,Cn* ,S and *Ci* are syntactic constituents
Semantic rule: *S0 ! F*(*C*'1,*C*'2….,*C*'*n*) ,S' and *C'I* are corresponding semantics.
Compositionality is a very important principle in analyzing semantics of a sentence. The process of identification of non-recursive cores of various types of phrases in natural language text is an intermediate step providing input to further full parsing stages. A number of methods for chunking have been explored for English & other European languages since [3] introduced a context free grammar based approach for chunking of English sentences.

However, for the languages like Kashmiri it is quite challenging like in most free world languages. The internal structure of Kashmiri is often in an unrestricted order as is often the case with verb chunks. Here chunk identification becomes quite difficult. Kashmir language has lack of annotated corpora of language text. To overcome this we need rules followed by manual validation.

III. Resources Needed For Semantic Parsing

To parse any language the basic knowledge needed is about words and the relationship between the words. For English language we have resources like WORDNET-a large lexical database containing rich information about words and concepts. This level of knowledge is known as word level knowledge .Next higher level is FRAMENET-a resource of information

about different situations. This is also known as Sentence level knowledge. In addition to these we also need manually defined rules which have mapping from syntactic structures to semantic relations.

In order to classify a language with a significant performance a large annotated corpora is required. But such corpora is unavailable in Kashmiri language so we need to manually atomize the annotation by writing some generalized rules.

A. Frame net And Word net

Frame Net specifies a great deal of structural information both within and among frames. FrameNet data is meant to be lexicographically relevant, not statistically representative, as in [6] and it can be used as a starting point for parsing. To build Rules we have to extract syntactic features from the frame net and some features can be extracted from sentence context.

Kashmiri is a V2(verb second) language like German, the tensed clauses are subject to verb second constraint; the finite verbal element in these clauses always occurs in the second position i.e., the position immediately following the first phrasal constituent, as in[4]. We need to identify head words (content words) and non-head words (function words)

After all the syntactic features have been extracted and identified we need to apply semantic roles to create a mapping from syntactic features to semantic roles.

Aslam chu kita: b para: n
Aslam is book read –p
Aslam is reading a book

s>cha asa: n a: sa: n
She is laugh-imp be –i
She laughs frequently

Rahim d'ut nasi:mas kalam
Rahim -erg gave nasim -dat pen
Rahim gave a pen to nasim

Feature sets are arranged in a list the order of which is identical to that in the sentence. There can be multiple annotations for each frame to demonstrate different syntactic realizations. All these are collected and stored in the list for that frame which should also contain target word ,its syntactic category and name of the frame. All the frames in the frameset have to be arranged in this form to be used by semantic parser.

The WordNet is used to identify semantic features that can be attached to lexical units like attribute relation, adjective/adverb classifications and then they are attached together with words so that they can be used for parsing.

B. Determining Semantic Relations

1) ***S**yntax:* The syntax provides us with most of the clues about the semantic structure of a sentence. According to the principle of semantic compositionality, a syntactic constituent carries a relatively complete and independent unit of semantics, so the syntactic parsing into smaller syntactic constituents is necessary. In the other hand the syntactic relation can imply, and in many cases directly identify, semantic relations. Although for efficiency consideration, the number of syntactic rules has to be small in order to make the language easy to use, it is still the major tools for human to figure out or establish semantics structures.
2) ***L**exicon:* The lexicon is divided into content words and function words based on their syntactic functions. Content words carry semantics representing facts, concepts and events. Content words include nouns, verbs, adjectives and adverbs. Content words are open, and new words are being created constantly. The semantics of individual content words will not be handled by the semantic parser, because the parser is used to analyze the structure of the semantics, but not the semantics itself. It should be noticed that morphological change of a content word may constitute additional semantics and the relation between its original semantics and the additional semantics should be studied. For example "bird" is the plural form of the word "birds". The morphological change from \bird" to \birds" provides the quantity. This additional semantics and the original semantics of \bird" form an attribute modification relation. All words other than content words belong to function words, such as preposition, determiners, auxiliary verbs and etc. Their major function is to help identify syntactic and semantic relations. Some of them carry semantics while some of them don't. Function words play important roles in composing semantics in content words and they should be paid close attention. The lexicon is classified into several categories based on words syntactic features. But the number of categories is relatively small due to the simplicity of syntax that they do not reveal much semantic information.
3) ***O**ntology:* The syntactic relations and semantic relations do not have a one-to-one mapping, and there are usually a smaller number of syntactic relations, the same syntactic relation may represent different semantic relations in different situations. So the situation needs to be considered in order to figure out the semantic realization of the syntactic relations. "Situation" is an ontology level concept and we need a shallow level ontology database that can be used to identify, distinguish situations and define how semantic relations are realized by syntactic relations in these situations. The

solution to this used in the semantic parser is FrameNet. The frame semantics thinks that a sentence tells a situation, denoted as “frame", and various syntactic constituents in the sentence play semantic roles in the situation, denoted as “frame elements". These roles indicate the semantic relations in these syntactic constituents in the sentence. For example “I paid you 30 rupees for the book" tells a commercial transaction situation (frame) according to the FrameNet. The semantic role “Agent" is assigned to “I", the subject of the sentence, indicating who initiates the transaction. “Paid" is the action verb that triggers the situation (frame). The role “Money" is assigned to “30 rupees", object of the verb “paid", as the expense for the “Agent" in the transaction. The “BOOK" is the “goods" role in this situation (frame). FrameNet has defined most frequently used frames to represent various situations and most importantly it defines in each frame, how frame elements are syntactically realized. This explains why the syntactic relation can represent different semantic relation and gives us a way to find it out. But it should be noticed that FrameNet is a shallow level ontology. It does not contain such high level knowledge that a ball is round or the sun rises in the morning. It is very useful for us to analyze the semantic structure of a sentence, but it is very linguistics oriented and its ontology should normally be applied in one sentence and hence it cannot be used to analyze contextual meaning which need very high level ontology and common sense. In addition to knowledge about situations and what semantic relations the syntactic relations play in these situations, knowledge about word relations is also necessary. According to the principle of compositionality, the semantics of the whole sentence can be decomposed to the smallest semantic units. To figure out how word semantics can be composed into bigger semantic constituents, we need knowledge about their semantic relations. The semantic relations among words can be within the same category or across categories. The WordNet indexes content words based on their senses rather than in an alphabetical order. According to the word sense, the WordNet connect semantically related words through links indicating some semantic relation. The link can be within the same category or across categories and it represent various semantic relations such as “synonym",”antonym",”part-whole" relation, “Isa" relation and etc. Such information about semantic relations between words enable us to find out how word semantics can forms bigger semantic constituents. We can say FrameNet provides sentence level ontology, and then WordNet provides lexical level ontology.

IV. Semantic Parsing

The rule based semantic parser implements 3 steps:

1. Undergo syntactic parsing to transform into intermediate format.
2. The semantic role assignment.
3. Application of default rules.

A. Syntactic/Semantic Analyzer

Semantic parser works on the dependencies between the words identified after syntactic analyzer. The intermediate format is generated where target words and syntactic arguments are identified and which are then matched against the rules The intermediate format also identifies the word level semantics and relations that have direct syntactic correspondence. It also identifies sentence types like assertion, query, and command.

Another important feature of analyzer is that it is based on feature of augmented grammar ,that has the ability of detecting the grammatically incorrect sentences and subsequently grammatically incorrect sentences will be rejected.

Kita:b khay:em
I ate the book

This sentence is semantically wrong but syntactically correct.

B. Semantic Role Assignment

Semantic role assignment starts with the identification of frames associated with the target word. Then the most appropriate match among all rules is selected and correct frames are identified and semantic roles are assigned. Here we can employ a scoring function to evaluate similarities between syntactic features. The matching starts form right to left and when any match is found score is increased by 1. The searching is unidirectional and we cannot go back once a match has been found ,this ensures that syntactic order has been maintained. however we can move forward to left but we can’t go back to right. we can also have manual selection for multiple possible syntactic realizations. We match the intermediate form with all the rules available for the word and compare the scores and the rule with the highest score is selected. This is done recursively till all the roles are assigned to all the target words.

C. Applying Default Rules

Since Fragment does not define roles for all the elements and meaning of some of the constituents cannot be defined , so to handle such constituents and for complete semantic interpretation we have to define certain set of rules that can be generalized and applied at the end of semantic parsing.

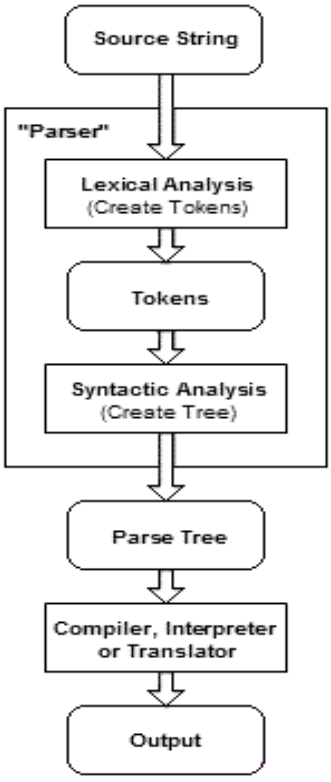

Figure 1: Generalized Flow Chart for Parsing

V. Related Work

Similar approach has been successfully applied for English language. All previous work in semantic parsing has exclusively focused on labeling semantic roles, rather than analyzing the full structure of sentence semantics, and is usually based on Statistical models -e.g. as in [2] and [8].

VI. Conclusion

We have described a rule based approach of semantic parsing for kashmiri language .This approach is closer to the ways human interpret the language .This approach analyses the constituent of a sentence and analyze the meaning of entire sentence on the basis of smaller semantic units. This semantic parser relies on the rules retrieved from the frameset and WordNet and since Kashmiri language lacks a complete lexical database, so we need to manually add rules to it.

References

[1] Lei Shi and Rada Mihalcea *"Open Text Semantic Parsing Using FrameNet and WordNet"* University of North Texas Department of Computer Science and Engineering and University of North Texas

[2] Gildea and D. Jurafsky. 2000. "*Automatic labeling of semantic roles".* In Proceedings of the 38th Annual Conference of the Association for Computational Linguistics (ACL-00), pages 512520, Hong Kong, October.

[3] P. Abney. Parsing by chunks. In Robert C. Berwick, Steven P. Abney, and Carol Tenny, editors, "*Principle-Based Parsing: Computation and Psycholinguistics"*, pages 257–278. Kluwer, Dordrecht, 1991.

[4] Bharati, R. Sangal, D.M. Sharma, and L. Bai 2006. Anncorra*:"Annotating corpora guidelines for pos and chunk annotation for Indian languages".* LTRC- TR31.

[5] M. Bhatt.1999."*Verb movement and the syntax of Kashmiri"*, volume 46. Springer Netherlands.

[6] Johnson, C. Fillmore, M. Petruck, C. Baker, M. Ellsworth, J. Ruppenhofer, and E. Wood. 2002*."FrameNet: Theory and Practice".* http://www.icsi.berkeley.edu/ framenet.

[7] Miller. 1995."Wordnet: A lexical database. Communication of the ACM*"* 38(11):39–41.

[8] Fleischman, N. Kwon, and E. Hovy. 2003. Maximum entropy models for FrameNet classification. "In *Proceedings of 2003 Conference on Empirical Methods in Natural Language Processing"* EMNLP-2003, Sapporo, Japan.

[9] Hook, Peter Edwin and Omkar N Koul 1984. "*On the Grammar of Derived Transitives and Causatives in Kashmiri".* In Koul andHook (Eds.) 1984.

[10] Luke S. Zettlemoyer and Michael Collins. 2005. *"Learning to map sentences to logical form: Structured classification with probabilistic categorical grammars"*. In Proceedings of the 21st Conference on Uncertainty in Artificial Intelligence.

[11] Li Shiuan and C. Ting Hian Ann. 1996. *A "Divide-and- Conquer Strategy for Parsing"*. In Proceedings of IWPT 1996. 57–66. Park, S. B. and Zhang, B. T.: Co-trained support "*vector machines for large scale unstructured document classification using unlabelled data and syntactic information*". Journal of Information Processing and Management (2004) 40: 421-439.

[12] PVS and G. Karthik. 2007. "*Part-of-speech tagging and chunking using conditional random fields and transformation based learning. Shallow Parsing for South Asian Languages"*, page 21.

[13] Kachru, Braj B. 1973. *"An Introduction to Spoken Kashmiri.Urbana":* University of Illinois.

[14] Koul, Omkar N and Kashi Wali *"Modern Kashmiri Grammar".* Springfield: Dunwoody Press.

