Semantic Parsing of Tamil Sentences

Balaji J, Geetha T V, Ranjani Parthasarathi
Dept of CSE & IST,
Anna University, Chennai – 600 025
jkb.8in@gmail.com, tv_g@hotmail.com,
ranjani.parthasarathi@gmail.com

ABSTRACT

In this paper, we propose a rule-based approach for the identification of semantic sub-graphs from Tamil sentences. In order to achieve the goal of semantic sub-graph identification and construction, we use a semantic graph based representation called Universal Networking Language (UNL), which is a directed acyclic graph representation. To identify and build the semantic sub-graphs, we classify the rules based on morpho-semantic features which include word associated features and context based features. The rules are performed in two stages; one while building the simple UNL graphs and one after the simple UNL graphs construction. We have identified 18 rules for sub-graph identification and construction.

Keywords: Nested graphs, Universal Networking Language, Semantic Graph representation

Proceedings of the Workshop on Machine Translation and Parsing in Indian Languages (MTPIL-2012), pages 15–22,
COLING 2012, Mumbai, December 2012.

1. Introduction

Semantic interpretation and representation of natural language texts is an important task of natural language processing. A number of semantic representations have been used to represent natural language using conceptual representations. Graph based semantic representations such as semantic networks (Masterman, 1961) is a declarative graphic representation consisting of definitional, assertional, implicational, executable, learning and hybrid networks. Silvio (1961) proposed correlational nets based on the relations such as part-whole, case relations, instance, subtype and dependence relations. Yet another graph representation that Hays (1964) proposed dependency graphs based on minimal syntactic units and conceptual graphs (Sowa, 1976) represents relations using inferences of first order logic. Similar semantic graph representation which contains semantic relations and attributes is Universal Networking Language (UNL) relations (UNDL, 1997) consists of 46 relations include agent, object, place, time, conjunction, disjunction, co-occurrence, content, quantity etc. represent semantics of natural language. Our work focuses on the identification of sub-graphs of semantic graphs for which we use UNL representation. The detailed study on UNL is described in section 3. Identifying sub-graphs from a semantic graph though a difficult task is necessary to obtain the boundaries between complex semantic entities and which in turn helps in understanding the complete meaning conveyed by a natural language sentence.

In this paper, we focus on identifying sub-graphs and building a nested graph structure for different types of natural language sentences in morphologically rich and relatively free word order languages. In this paper, we describe the identification of semantic sub-graphs from Tamil, a morphologically rich and relatively free word order language. Here, we define two set of rules one based on word associated features and another based on context oriented features to build a hyper-graph or nested UNL graph (NG) structure to represent sub-graphs of different phrases and clauses of sentences. We build the nested graph structure in two stages; one, the identification of sub-graphs while building simple graphs (SG) and second, the identification of sub-graphs after the simple graph (SG) construction. The paper is organized as follows. Section 2 discusses the related works carried out using nested graph structure. Section 3 discusses the nested UNL graphs and the rules defined for sub-graph identification and construction. The evaluation and performance of the defined rules are investigated in section 4. Finally, section 5 concluded with future enhancements.

2. Related Work

Since we are focusing on the identification and construction of nested UNL graphs, first we discuss UNL in detail and the other similar semantic graph representations in this section. UNL (UNDL, 2011) is an electronic language designed to represent semantic data extracted from natural language texts. UNL consists of Universal words (UWs) represents concepts, relations represent the semantic relationship between concepts and attributes represents mood, aspect, tense etc. UNL representation is a directed acyclic graph representation in which nodes are concepts and links are relations exist between the concepts.

(Blanc, 2000) described the French UNL Deconverter in which the issues in representing the semantic characteristics of predicative concepts have been discussed. Dikonov (2008) discussed the representation of UNL graphs by segmenting complex graphs into simple graphs by applying rules based on propositions. Coreferential links are also considered in segmenting the UNL

graphs. (Jain &Damani, 2008) described the identification of scopes by relative positions in a phrase structure tree. The author classified the relations into cumulative and others. In the same vein, this paper also focuses on the identification of sub-graphs of UNL semantic graphs.

Similar to the graph based semantic representation discussed above Chein et al (1997) presented a general framework for nested graphs provided with morphism which is a mapping between two graphs that induces reflexive and transitive relations for defining nested conceptual graphs in preorder. The simple conceptual graphs have been generalized by reasoning of objects based on projection operation. A survey on frequent sub-graph discovery has been described by (Krishna et al, 2011) in which the popular graph mining algorithms have been compared. The author also discussed the essential factors of various graph algorithms for discovering the sub-graphs.

In this paper, we focus on the identification and construction of semantic UNL nested graphs from Tamil sentences. Unlike other sub-graph identifications for languages such as French and English, we use the word associated features and context based features such as UNL semantic relations instead of parsing the sentences. However, the other approaches for UNL nested graph identification explored some UNL relations, we also explore more UNL relations in the sub-graph identification.

3. Nested UNL Graphs representation

The UNL representation is said to be a hyper-graph, when it consist of several interlinked or subordinate sub-graphs. These sub-graphs are represented as hyper-nodes and correspond to the concept of dependent (subordinate) clauses, and a predicate. They are used to define the boundaries between complex semantic entities being represented. A scope is a group of relations between nodes that behave as a single semantic entity in a UNL graph. For instance, in the sentence "John killed Mary when Peter arrived", the dependent clause "when Peter arrived" describes the argument of a time relation and, therefore, should be represented as a hyper-node (i.e., as a sub-graph) as represented below:

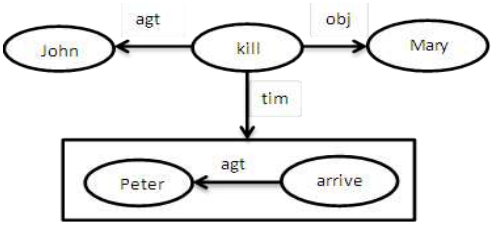

FIGURE 1 UNL graph representation with nested sub-graphs for "John killed Mary when Peter arrived"

Scopes are used to segregate subordinate clauses such as adverbial clauses, adjectival clauses and nominal clauses. Scopes focus on identifying relations that exist between different types of subordinate clauses and does not focus on the relations exist between words. Sub-graphs are identified only when the semantic unity is formed by the interlinked nodes and are semantically ambiguous.

A hypergraph can be defined as a generalization of a graph in which an edge can connect to any number of vertices. The set of vertices of a hypergraph that is connected by an edge is normally associated in some way and can be considered as sub-graphs. Hypergraphs can be formed when nodes of one sub-graph edge may originate/ terminate from/to a sub-graph considered as a single entity. Therefore, hypergraph has better expressive power than an ordinary graph. UNL representation is a directed acyclic graph representation in which the set of nodes interlinked by edges can represent a single semantic entity which is formally defined by UNDL as scopes (or) hyper-nodes (or) hyper-graphs as mentioned earlier.

Let the graph G = {N_1, N_2 …… N_i} where the set {N_1, N_2 …… N_i} consists of nodes connected by relations {R_1, R_2 ….. R_j}. From graph G, the hidden nested graphs (NG) are identified by a set of rules. The rules are based on morphological suffixes, POS and semantic information associated with word and the context. The pseudo code is for rules are given in Figure 2.

3.1. Rules for Nested Graph Identification

Balaji (2011) presented a rule-based approach for building the UNL graphs of Tamil sentences using morpho-semantic features. The 53 rules defined in their approach were utilized for converting Tamil sentences into simple UNL graphs. In this paper, we incorporate a new set of 18 rules for sub-graph identification with the existing set of 53 rules originally proposed by Balaji (2011). We explore new rules based on word and context based features for the identification of sub-graphs during two stages one while constructing simple UNL graphs and one after simple UNL graph construction has been completed. The rules defined by us are categorized based on two stages where they are used and have been listed below.

The rules are performed in two stages.

> Stage 1: identify sub-graphs while constructing simple UNL graphs (SG) using features such as lexical and word based information conveying semantic relations between nodes in SGs. Sub-graphs that fall under this category are the different phrase types.
>
> Stage 2: identify sub-graphs (NG) after the construction of simple graphs (SG) by preserving the context level information in the nodes and the relations connected between the nodes of SGs. Sub-graphs that fall under this category are the different clause types.

As discussed earlier, rules are classified based on word associated features such as morphological suffix, POS, UNL semantic constraints and in certain cases the word itself which may convey UNL semantic relations, and the context based features such as UNL semantic relations. These rules set are utilized in the identification and construction of nested sub-graphs.

1) Let the nodes N_i and $N_j$where i~=j, be connected with UNL relations (R) such as and, mod, pos. Then the UNL graph R(N_i, N_j) can be a sub-graph (NG). In the UNL representation, hyper-nodes are indexed by ":XX", where XX is a two-digit hyper-node index called scope-id.

Example 1:azakiya poonga (beautiful park). This example has the node N_i as "azakiya" and node N_j as "poonga". The nodes N_i and N_j are connected by "mod" relation and marked with scope-id as nested graph (NG). In Fig 3, the scope identifier is assigned to the headword "poonga" of simple UNL graph. Similarly, the nested graphs are identified using other relations.

Let the Graph (G) be represented as G (V, E) where V-Vertices and E-Edges
Notations:
Simple UNL graph – SG, Nested UNL sub-graph – NSG, Scope Identifier - SCP_{id}
Morphological suffix – MS
Parts of Speech – POS
UNL Relations – UNLR
Rules – R
Input: Natural Language Sentence
Output: Nested UNL Graphs
Stage1:
Refer (Balaji et al, 2011) for building simple UNL graphs.
If(R (MS) ||R (POS)) {
(POS ϵ Adjective, Adjectival Noun, MS ϵ Adjectival Suffix, Genitive case)
Set SCP_{id} for SG;
NSG ← SCP_{id} (SG); (SG consists of Concept-Relation-Concept (i.e. V-E-V))
} else if (R (Connectives)) {
(Connectives ϵ Postpositions, Conjunctions)
Set SCP_{id} for SG;
NSG ← SCP_{id} (SG);
}
Stage2:
if (UNLR (modifier || conjunction || possessor)) {
Set SCP_{id} for SG;
NSG ← SCP_{id} (SG);
}

FIGURE 2 Pseudo code for Nested UNL Graph Identification and Construction

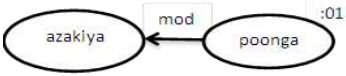

FIGURE 3 Simple UNL graph representation for "azakiya poonga" with scope id (:01) assigned to the headword of the graph

2) Rule set defined above is to identify simple graphs that can be a sub-graph (i.e. node connected to another node can be a sub-graph) as described with an example. The next rule is to identify the nested graphs of types

a) node (N) connected to a sub-graph (SG) and vice versa N→SG (or) SG→N
b) Sub-graph(SG) connected to another sub-graph (SG) SG→SG^1

a) N→SG (or) SG→N

Example 2: azakiya poongavil sandhiththaan (met in a beautiful park). This example clearly illustrates the need of nested graphs. The verb "sandhi" is connected to "poonga" by "plc" relation. The graph obtained with "plc" relation gives only the partial meaning of the sentence. In order to obtain the complete meaning of the sentence in a graph, we can represent in two ways. One is by simply connecting the nodes with the appropriate UNL relations in a sentence and another is by representing the sentence in which shows the exact meaning of the sentence. Both Fig 4 and 5 shows simple graph and nested graph representations respectively.

FIGURE 4 Simple Graph representation of "azakiya poongavil sandhiththaan"

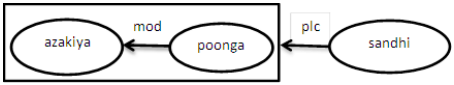

FIGURE 5 Nested Graph representation of "azakiya poongavil sandhiththaan"

b) SG→ SG[1]

A sub-graph is connected to another sub-graph when the UNL relations conveyed by the lexical endings such as "aal" and some natural language words, normally referred as connectives (mostly postpositions and conjunctions in Tamil) such as "maRRum", "paRRi", "enpathu" etc. Different clauses such as adverbial, adjectival and nominal clauses are formed based on the connectives and the semantic sub-graphs are identified using these connectives.

Example3: azakiya poonga maRRum periya aaRu – beautiful park and big river

Figure 6 shows the sub-graph to sub-graph representation. When representing as a simple semantic graph, the headwords of both the sub-graphs "poonga" and "aaRu" are connected with the UNL relation "and" in which the word "maRRum" indicates "and" relation.

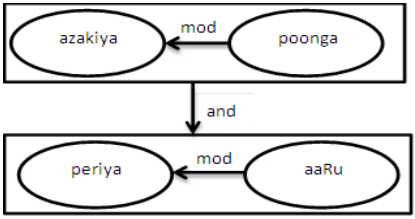

FIGURE 6 Nested graph representation of "azakiya poonga maRRum periya aaRu"

4. Evaluation

We investigate our rule-based approach which consists of 18 rules for identifying and constructing nested graphs using 500 sentences. Among the total of 500 sentences, 160 sentences contain adverbial clauses, 140 are adjectival clauses and 200 are nominal clauses. In addition to the different types of clauses, different phrase types are also taken into consideration for evaluating the set of rules. The output graphs are evaluated in an ad hoc manner. Table-1 shows the analysis and performance of nested rules for different types of sentences. Table-1 also shows the actual number of sub-graphs (in percentage) present under each category of sentences and the number of sub-graphs (in percentage) identified by our rules in each category of sentences. However, the rule-based approach produces better results, additional rules are required to improve the performance of our approach. Moreover, in order to identify the sub-graphs automatically and make it domain independent, we are also focusing on machine learning approaches for the identification and construction of nested UNL graphs.

Table-1: Performance Evaluation of Nested UNL Graphs

Category of sentences	Precision	Recall
Adverbial clauses	0.55	0.34
Adjectival clauses	0.66	0.37
Nominal clauses	0.64	0.45

The reason behind the low recall is the number of sub-graphs correctly identified is comparatively low when compared to the number of sub-graphs actually present in the corpus. The complex sentences can have the possibility to have incorrectly mapped concepts. This is because some rules may conflict each other. To improve the performance of our approach and to increase precision and recall, more number of disambiguation rules are needed.

5. Conclusion

In this paper, we described rules for identifying sub-graphs of UNL semantic graphs. The rules are classified into different categories – one is the identification of sub-graphs while constructing simple graphs and another is the identification of sub-graphs after simple graph construction. The features considered for defining the rules are word associated features such as morphological suffixes and POS, and context associated information such as UNL relations. The defined rules are tested with health domain corpus and it produces significantly better results. Further, we enhance the identification of sub-graphs using machine learning approach so as to achieve the task to be domain independent.

References

Balaji J, T V Geetha, Ranjani, and MadhanKarky, (2011), Morpho-semantic features for rule-based tamilenconversion. International Journal of Computer Applications, 26(6):11{18, July 2011. Published by Foundation of Computer Science, New York, USA

Blanc Etienne, (2000), From the UNL hypergraph to GETA's multilevel tree, MT 2000 Conference, Exeter, UK 11/2000

Ceccato, Silvio (1961) Linguistic Analysis and Programming for Mechanical Translation, Gordon and Breach, New York.

Chein, Michel and Mugnier, Marie-Laure, (1997), Positive Nested Conceptual Graphs, Proceedings of the Fifth International Conference on Conceptual Structures: Fulfilling Peirce's Dream, page 95--109

Dikonov V, (2008), UNL Graph Structure, Informacionnye process, vol. 8 #1

Hays, David G. (1964) "Dependency theory: a formalism and some observations," Language 40:4,511-525.

Jain, M. and Damani, O. P. (2008), English to UNL (Interlingua) Enconversion, Indian Institute of Technology, Bombay, India

Masterman, Margaret (1961), Semantic message detection for machine translation using an Interlingua, Proc. 1961 International Conf. on Machine Translation, 438-475.

Sowa, John F, (1976) "Conceptual graphs for a database interface," IBM Journal of Research and Development 20:4, 336-357.

UNDL, (2011), www.undl.org, accessed Sep. 2011

Varun Kirshna, Ranga Suri N N R, Athithan G, (2011), A Comparative survey of Algorithms for Frequent Sub-graph Discovery, Current Science, Vol. 100, No. 2, 25, JAN 2011