Generation of MRS Abstract Predicates from Paninian USR

| Sukhada | Veera Hymavathi Sirisipalli | Soma Paul |
|-------------------------|----------------------------------|------------------------|
| IIT (BHU), Varanasi | IIT (BHU), Varanasi | IIIT Hyderabad |
| sukhada.hss@itbhu.ac.in | hymavathisirisipalli56@gmail.com | <u>soma@iiit.ac.in</u> |

We are developing the architecture and rules of a transfer grammar module from Universal Semantic Representation (USR) to ERG meaning representations which is couched in Minimal Recursion Semantics (MRS; Copestake et al., 2005). USR is inspired from Indian Grammatical Tradition (IGT). USR is a language independent representation, it can ideally be used to generate any natural language. Currently some Indian languages such as Hindi, Tamil and Bangla generation are targeted from the USR. For English generation from the USR, we are using the open source ACE¹ generator. For this task, the USR is mapped to ERG meaning representation which the ACE generator uses as a resource grammar. Since, both USR and MRS are semantics based representations, we assume that the USR-MRS transfer would be straightforward. While designing the transfer module from USR to ERG based MRS representation, we have come across various abstract predicates as described in ErgSemantics_Basic². These abstract predicates are used to represent the semantic contribution of grammatical constructions or more specialized lexical entries such as compounding or the comparative use of *more*.

This paper presents the strategy and implementation for postulating the abstract predicates from the information given in USR. Before describing the strategy, we will examine the information encoding system of USR and also various abstract predicates as postulated in ERG.

Universal Semantic Representation following Indian Grammatical Tradition

The Universal Semantic Representation (USR) is a csv formatted multilayered information packaging system that encapsulates lexico-conceptual, intrasentential and discourse level information. The uniqueness of this representation is that information of each layer is distinctly yet interactively maintained through attribute value matrix and co-referencing as shown in sentence (1). The USR for the semantics of sentence (1) is given in Table-1:

(1) hari ne apane guru jī ko garama dūdha aura miţhāī dekara ābhāra vyakta kiyā hari ERG his teacher RESPECT DAT hot milk and sweet offering gratitude express do.PST 'Hari expressed gratitude to his teacher by offering hot milk and sweet'

| Concept | hari | apanā/ he | guru/ teacher | garama /hot | dūdha/ milk | mițhāī/ sweet | de/ offer | ābhāra /gratitude | vyakta+kara/ex press-yA/past |
|----------------------|-------------------|--------------|------------------|----------------|-------------------|------------------|--------------------|----------------------|---------------------------------|
| Index | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| Semantic Category | person | | animate | | mass | | | abstract | |
| Morpho-semantics | [m sg a] | | [m sg a] | | [- sg a] | [- sg a] | | | |
| Dependency | 9:k1 ³ | 3:r64 | 7:k4⁵ | 5:mod | 7:k2 ⁶ | 7:k2 | 9:rpk ⁷ | 9:k2 ⁸ | 0:main verb |
| Discourse | | 1:coref | | | | | | | |
| Speaker-view | | | respect | | | | | | |
| Construction | conjunction:[5,6] | | | | | | | | |
| Sent Type | affirmative | | | | | | | | |

Table-1: USR for the sentence (1)

⁵ k4: recipient

- ⁷ rpk: sequential
- ⁸ k2: theme

¹https://pydelphin.readthedocs.io/en/latest/guides/ace.html

² https://github.com/delph-in/docs/wiki/ErgSemantics Basics

³ k1: doer/agent

^₄ r6: genitive

⁶ k2: object

The 1st row represents concepts (not words) which are represented in terms of equivalents from different languages (in this case Hindi/English). The 3rd and 4th rows capture the semantic category, inherent gender-number-person and other semantic information such as comparison, causation which can be marked in languages morphologically. Relations among concepts are captured in terms of dependency relation in the 5th row. The digits are the index number of the head and the relation with the head is stated after the colon. Discourse level information such as inter-sentential information, co-referencing are conveyed in the 6th row. In the 7th row, the intention/speaker's view is captured. For instance, the speaker addresses the teacher with respect (in Hindi, the corresponding word is $j\bar{i}$, in Bangla $b\bar{a}bu$, in Telugu $g\bar{a}ru$ and so on). Conjunction is represented in the Construction row. Finally the sentence type is also specified.

Abstract Predicates

ERG has around 108 Abstract predicates. They can be classified into broad categories as:

- a. Quantifier
- b. Abstraction
 - i. Comparison
 - ii. Generic
- c. Construction
- d. Other

Under each category, following abstract predicates are handled in our transfer grammar so far:

| а. | Quantifier | | |
|----|--|---|---|
| | > def_explicit_q > def_implicit_q > every a | number_q pronoun_q proper_q | ≻ udef_q ≻ which_q |
| b. | Abstraction i. Comparison ≻ comp ≻ comp_equal | p p 4 | ≻ comp_less ≻ super |
| | ii. Generic ≻ card ≻ dofw > generic_entity > loc_nonsp | measure mofy named ord | > person > place_n > pron > thing > time_n |
| C. | Construction ≻ compound ≻ implicit_conj | ≻ parg_d> recip_pro | > subord > neg > poss |
| d. | Other ≻ property | > prpstn_to_prop | > unknown |

In this paper, we attempt to identify where and how information is encoded in USR which enables us to postulate the aforementioned abstract predicates. In most of the cases, semantic information encoded in USR are used to determine the abstract predicate while there are few cases where we are currently using mainly lexical information to postulate abstract predicates.

Quantification

As described in the ErgSemantics_Basic document, the ERG assumes that all instance variables (of type x) are bound by a generalized quantifier. Such an assumption is not taken in USR. In order to predict the right quantification predicate, we use the following USR information:

| MRS quantifier | Context | USR information utilized | Example |
|-------------------|-----------------------------------|---|--|
| proper_q | When the noun is a proper noun | Named entities are annotated as sem_cat | Hari - sem_cat: person London - sem_cat: place |
| pronoun_q | All personal pronouns | Speaker, addressee are annotated for 1st and 2nd person. 3rd person pronouns are co-referred in the larger discourse | I/we < speaker+number You < addressee 3rd person pronouns < co-reference+gender+number |
| which_q | All wh questions | The concepts for any wh pronoun is represented as <i>kim</i> in the concept_row with the appropriate dependency relation marked in the dep rel row. | Who < <i>kim</i> +kartā (doer/agent) When < <i>kim</i> +kālādhikaraņa (temporal) Where < <i>kim</i> +deśādhikaraņa (spatial) |
| def_explicit_q | Possessive nouns and pronouns | Possessive information is annotated in terms of r6 relation between the head and modifier | Ram's book < book Ram:r6 |
| def_implicit_q | Spatial and temporal adverbs | Concepts from concept row are the triggers | yahāṃ (here), vahāṃ (there), aba (now), āja (today), kala (tomorrow) |

Comparison

USR encodes the information about comparison_equal and comparison between the two and among many at two different places: (a) at Morpho-semantic layer as shown in the table below and (b) at Dependency layer as a relation between the two nominal entities: one that is compared with and the other being compared through the relation is *ru* (*relation upamāna*). Here is an example:

| Concept | mīrā | sītā | sumdara/beautiful | hai/be |
|-------------------|---------|--------------------------------------|--|--------|
| Index | 1 | 2 | 3 | 4 |
| Semantic category | person | person | | |
| Morpho semantics | | | comper_more | |
| Dependency | 4:kartā | 1: <i>upamāna</i> 'compared with' | 4:kartā_samānādhikaraṇa 'miraness and beauty resides in the same locus' | main |

The English sentence generated from this USR would be "Mira is more beautiful than Sita"

In MRS, one predicate for comparative is postulated as "*comp*". It is a dyadic predicate with the verb being the first argument and Sita (who is compared with) being the other argument. The semantic analysis of the two frameworks is, however, different even though mapping could be done from USR and MRS and the expected English sentence could be produced by our system.

Abstraction

This category consists of cases where MRS representation goes one level more abstract than the surface predicates to capture certain generalization in the representation. For example, instead of specifying days of week or months of a year by its name or year in a calendar, the representation postulates abstract predicates *dofw, mofy* and *yofc* respectively. Similarly *card* predicate is postulated for numbers and numerical modifiers, *pron* predicate for personal pronouns and so on. In the paper, other predicates under abstraction will be discussed in detail with examples.

Construction

This category includes what we commonly call as construction - form-meaning pairs. For different constructions different kinds of information from USR is being utilized for the mapping. For instance, multi-word compounds are currently represented with a + (plus) symbol such as basa+adda ('bus stand'). Genitive information (used for the postulation of *poss*) is specified through a dependency relation between the head noun and its genitive modifier. Tense-aspect-modality (TAM) information and sentence type has been used to postulate passive predicate.

Implementation of Transfer Module in CLIPS

The implementation is done at two levels: (a) Determining the abstract predicate (b) Specifying the feature structure description of the abstract predicate. For (a), information from USR has been utilized as discussed in the previous section. Once the abstract predicates are identified, we run a sentence with the target abstract predicate in ACE parser, find out the feature structure description and add it in the dictionary along with the abstract type if it is not already present. Thus the lexicon for abstract predicate is created. During English sentence generation from USR via MRS, this dictionary is consulted for framing the appropriate MRS for a given USR which in turn is used by ACE generator as an input and the English sentence is generated. The postulation of abstract predicates from USR is executed in CLIPS (Giarratano, J. C. 1993).



Flowchart: Transfer Module with abstract predicates handling highlighted

Statistical Observation on Transfer Rules

Implementation of rules for creating the abstract predicates include three types of mapping: (1) Direct Mapping: A relation or a lexical concept from USR is directly mapped to MRS abstract predicate; (2) Indirect Mapping: Information encoded at multiple positions in USR are used to postulate the abstract predicate; (3) Constraint based mapping where the rule includes constraints to prevent wrong or overgeneration of abstract predicates. Table-2 gives the statistics of the three rule types.

We observe that rules written for Indirect mapping is the highest in number. Thus we conclude that information used for postulating MRS abstract predicates is distributed at different layers of the USRs.

| Rule | Number of rules | In percentage |
|--------------------------|-----------------|---------------|
| Direct mapping | 10 | 27.78% |
| Indirect mapping | 15 | 41.67% |
| Constraint based mapping | 11 | 30.55% |

Table-2: Statistical observation on transfer rules in CLIPS

Experimental Setup

We first created 5 sample sentences for each abstract predicate under consideration. These sentences were shared with a Hindi native speaker with those words for which MRS requires abstract predicate highlighted. The native speaker created 10 sentences with the highlighted words. Thus we created 262 test suite sentences for which annotators developed USRs. The MRS generator of which Abstract Generator module is a part was run on the 262 USRs.

Result and Error Analysis :

| Total no. | Total no. of expected | Abstract Predicates | Error Analysis | |
|-----------|-----------------------|---------------------|--------------------|-------------------------|
| | | gonoratou | Concept missing | Typographic al error |
| 262 491 | 401 | 469 | 16 | 6 |
| | 491 | 95.5% | 3.25% | 1.22% |

 Table-3: Result and Error Analysis for Abstract Predicates

References

Flickinger, Dan, Bender, Emily M. and Oepen, Stephan. 2014. Towards an Encyclopedia of Compositional Semantics: Documenting the Interface of the English Resource Grammar. Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014).

Flickinger, Dan, Bender, Emily M. and Oepen, Stephan. 2014. Towards an Encyclopedia of Compositional Semantics: Documenting the Interface of the English Resource Grammar. Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014).

Bharati, A., & Kulkarni, A. (2008). Information Coding in a language: Some insights from Paninian Grammar. Dhimahi, Journal of Chinmaya International Foundation Shodha Sansthan I (1), 77–91

Copestake, A., Flickinger, D., Pollard, C. et al. Minimal Recursion Semantics: An Introduction. Research Language

Computation 3, 281–332 (2005). https://doi.org/10.1007/s11168-006-6327-9

Giarratano, J. C. (1993). CLIPS User's guide. NASA Technical Report, Lyndon B Johnson Center .

Kiparsky, P. (1995). Paninian Linguistics. In Concise History of the Language Sciences (pp. 59-65). Pergamon.

Kulkarni, A. (2008). Panini's Ashtadhyayi A Computer Scientist's viewpoint, 4th APCAP NIAS Bangalore