Evaluation of Universal Semantic Representation (USR)

Kirti Garg
IIIT Hyderabad
kirti.garg@gmail.com

Soma Paul
IIIT Hyderabad
soma@iiit.ac.in

Sukhada
IIT(BHU), Varanasi
sukhada.hss@iitbhu.ac.in

Riya Kumari
IIT(BHU), Varanasi
riyatomar912@gmail.com

Fatema Bawahir
IIIT Hyderabad
bawahir.fatema@gmail.com

Abstract

Universal Semantic Representation (USR) is designed as a language-independent information packaging system that captures information at three levels: (a) Lexico-conceptual, (b) Syntactico-Semantic, and (c) Discourse. Unlike other representations that mainly encode predicates and their argument structures, our proposed representation captures the speaker's vivakṣā - how the speaker views the activity. The idea of "speaker's vivakṣā" is inspired by Indian Grammatical Tradition. There can be some amount of idiosyncrasy of the speaker in the annotation since it is the speaker's viewpoint that has been captured in the annotation. Hence the evaluation metrics of such resources need to be also thought through from scratch. This paper presents an extensive evaluation procedure of this semantic representation from two perspectives (a) Inter-Annotator Agreement and (b) Utility for downstream task of multilingual Natural Language Generation. We also qualitatively evaluate the experience of natural language generation by manual parsing of USR, in order to understand the readability of USR. We have achieved above 80% Inter-Annotator Agreement for USR annotations and above 80% semantic similarity in multi-lingual generation tasks suggesting reliability of USR annotations and utility for multi-lingual generations. The qualitative evaluation also suggests high readability and hence utility of USR as a semantic representation.

1 Introduction

Semantic Representations (SemRep henceforth) generally encode predicate-argument structure of a verb (Propbank (Kingsbury and Palmer, 2002) and Palmer(OnlinePalmer et al., 2005), FrameNet (Baker et al., 1998) along with some other grammatical information ranging from lexico-syntactic level information such as tense-number-person (AMR (Banarescu et al., 2013), MRS (Copestake et al., 2005) to discourse level information such as topic-focus, co-referencing and discourse relations (PDT (Sgall et al., 1992) (Böhmová et al., 2003), UCCA (Abend and Rappoport, 2013). However, no semantic representation, that we are aware of attempts to capture what we term as the speaker's vivakṣā - how the speaker views the activity. We design a Universal Semantic Representation (USR) that encodes "speaker's vivakṣā". The idea is inspired from the Indian Grammatical Tradition (IGT henceforth). IGT views languages as a holistic phenomenon. Words are not derived as isolated units in Pāṇini's grammar, but as units that are semantically connected with other words in the sentence (Raster, 2015). Sentences are connected across the discourse. This is explicitly recognized by the Paninian rule (A 2.1.1) : *samarthaḥ padavidhiḥ*[1]. Keeping in tune with IGT, USR is designed as a representation that encodes information ranging from lexico-conceptual to discourse level in a connected structural format. Since this is a very new kind of representation, the evaluation of such a resource also requires special attention.

This paper presents the evaluation metrics of USR from two perspectives: (a) the Reliability of USRs (b) the utility of USR in the domain of multi-lingual generation. Sentences are generated in multiple languages to check the universality of information encoded in USRs. We use simple reliable measures to evaluate and understand these properties.

The **quantitative evaluation** metrics are presented from two perspectives: (a) the Reliability of USRs, (b) the utility of USR. The utility is evaluated for multilingual generation measured through Semantic Textual Similarity (STS). We use simple

[1] An operation on words [takes effect only] when the words are semantically connected.

Proceedings of the 4th International Workshop on Designing Meaning Representations, pages 13–22
June, 2023. ©2023 Association for Computational Linguistics

reliable measures to evaluate and understand these properties.

The **qualitative evaluation** focuses on the usability of USRs in terms of readability of USR to generate natural language is examined. We also verify the adequacy of USR by manually generating natural language sentences from USRs.

A detailed analysis validates the proposed system as well as indicates areas of improvement. The feedback from these analyses is important for improving the information content and representation strategy of USRs.

Section 2 presents the design of USR. Section 3 studies Evaluation metrics in the context of other related works. Section 4 describes the quantitative evaluation metrics with results; while the qualitative measure is recorded in Section 5.

2 Design of USR

Unlike other representations that mainly encode predicates and their argument structures, the proposed representation captures the speaker's *vivakṣā*[2] - how the speaker views the activity. The idea of "speaker's *vivakṣā*" is inspired from Indian Grammatical Tradition (IGT henceforth). For example, how does the speaker's view differ when (s)he says 1 vis-a-vis 2? In Hindi, two different verb roots are used and the post-position on the seer also indicates different roles as shown in 1 and 2. In 1, Mira is an experiencer while in 2, the volitionality of Mira is maintained.

(1)	*mīrā ko*	*kala*	*cāṃda*	*dikhā*
	Mira.exprncr	yesterday	moon	see.int.pst

'Mira happened to see the moon yesterday'

(2)	*mīrā ne*	*kala*	*cāṃda*	*dekhā*
	Mira.kartā	yesterday	moon	see.tr.pst

'Mira saw the moon yesterday'

The activity of 'seeing' licenses[3] an animate *seer* and a *seen entity*. That is the *semantic frame* for the verb that every human being who knows the meaning of 'seeing' knows. But in communication, along with choosing the appropriate semantic frame, there occur two other important factors: (a) how the speaker conceptually cognizes the situation? (b) which linguistic expressions are available to translate that cognition into languages. For example, in the above examples, does the speaker want to express Mira's agency, or does (s)he want to foreground the appearance of the moon over the seer's agency? This is termed as the speaker's *vivakṣā*. Depending on that, the speaker would choose the best appropriate linguistic expressions to convey his/her thoughts. Our application task, namely Natural Language Generation (NLG) also motivates the requirement of capturing the speaker's *vivakṣā* in SemRep.

In order to generate a coherent and cohesive text, we require generative cues. Speaker's *vivakṣā* motivates those cues and we have decided to capture them in USRs through morphosemantics and dependency relations intra-sententially and also through discourse-level information.

USR encodes information at three basic levels: (a) Lexico-Conceptual (b) Syntactico-Semantic and (c) Discourse. This semantic information in USR is organized as features (in rows) and values, where the discourse relation and discourse co-referencing are accomplished through inter-USR linking which is established through Sentence_ID. Word_Index anchoring as shown in table 1. This representation is close to the Attribute Value matrix (AVM), but is easier to read and write manually, as well as process computationally.

Lexico-conceptual level: Conceptual Information which is generally expressed in terms of atomic words, multiword expressions or derived words are captured at this level. Currently, this level has information at 4 layers in USR. These layers (or rows) are (i) Concept row; (ii) Semantic Category; (iii) Morpho-semantic and (iv) speaker's view. Each entry to the concept row is an unambiguous representation of a concept. The ambiguity of a word is resolved in a very unique way in USR. Many SemReps use WordNet sense id as concepts. We propose to represent a concept in a multilingual set-up. For example, the lexeme in Hindi paḍha expresses two concepts: 'study' (as in *The boy studies in 7th standard*) and 'read' ('*the boy reads a book*'). This kind of ambiguity is handled at the conceptual level in the Concept Dictionary. This dictionary

[2] *śabdeṣvarthadānābhiprāyo vivakṣā* "vivakṣā is the intention of the speaker with regard to the meaning to be conveyed by the words" (Bhojaraja, 2007; Abhyankar, 1977). Abhyankar (1977) has also defined the term ”vaktur-vivakṣā”, in the same sense . As per "vivakṣātaḥ kārakāṇi" (Tripathi et al. 1986) kāraka roles in a sentence also apply according to the desire of the speaker.

[3] Selectional restriction of the verb which in IGT is known as a verb's *yogyatā*.

Concept Row	Sanskrit	Hindi	English	Bangla
paḍha_1	paṭha_1	paḍha_1	read_1	paṛa_1
paḍha_2	adhi+ī_1	paḍha_2	study_1	paṛa_2

Table 1: Concept Dictionary

has concept labels and equivalent concept labels in the languages under consideration. Currently, our lexicon has concepts in English, Hindi, Tamil and Bangla. The entry of a concept dictionary for the concept paḍha is the table 1.

USR has the Concept Label entry in the concept row which during generation selects concepts from the respective language cell depending on which language to be generated. In the current concept dictionary, there are 142037 labels for which Hindi and English concept labels are specified. For, 130948 concepts, Sanskrit labels are also attested in the dictionary. At the Lexico-conceptual level, the Semantic category row specifies the semantic category of a concept. Currently, four generic named entity categories are being annotated, namely- *per*(son), *org*(anisation), *place* and *other*. Apart from that, we mark *animacy* and *mass* categories.

Syntactico-Semantic level: Two types of relations capture information at this level: *kāraka* and *kāraketara* ('other than kāraka') (Kulkarni 2010) at the Dependency row. Pāṇini's system of knowledge representation is based on *kāraka* theory. There are six *kārakas* pointing out the relations between an event (or state) and its participants. They are *kartā*, *karma* (object), *karaṇa* (instrument), *sampradāna* (beneficiary), *apādāna* (source) and *adhikaraṇa* (time and location of action). *kāraketara* relations include relations between (a) noun and its modifiers; (b) verb and its verbal modifiers. There are a total of 42 dependency relations postulated till now in our work. **Discourse level:** Language as a mode of communication always occurs as a discourse in which a sentence generally has a connection or trace with the previous and following sentence. Discourse relations map such inter-sentential information which forms a coherent text. Co-reference is another discourse strategy to indicate two entities within a sentence or across sentences having the same referent. In USR, all intra-sentential discourse information is encoded in the Dependency row and inter-sentential discourse information is captured in the Discourse row. Currently, we are representing a few inter-sentential discourse relations as described in Das (2016) following IGT. They are *pratibandha* (If...then), *samānkāla* (when...then), *kāraṇa-kāryabhāva* (although), *hetu-hetumadabhāva* (because), *asāphalaya* (but), *anantarkālinatva* (then). More relations are being identified and a contrastive study with RST and PDTB tagsets are also being carried out. At present, if no explicit relation across USRs is marked, the default relation 'and' is presumed.

2.1 Example of USRs

Table-2 and Table-3 present examples of USRs that generate the discourse given in the following discourse 3.

(3) a. śāma ko eka yā do camakate tāre najara āte haiṃ.
'One or two shining stars come to our notice in the evening'

b. lekina kucha hī samaya meṃ unakī saṃkhyā baḍha jātī hai.
'But, within a short time, their numbers increase.'

Every sentence is given a unique sentence id. The first and second sentences are related with *asāphalaya* relation which is marked on the verb of the second sentence as Sentence_ID.Word_Index:Relation_name.

USR is designed to facilitate language generation tasks. USR is a text-based data structure and hence can be parsed both by the machine as well as humans effectively. The Sentence type row records the type of the sentence. Concepts specified in the Concept Row along with information from Morpho-semantic row, Semantic Category row determine the correct word forms. Speaker's View row information is used to postulate discourse particles that convey the speaker's view. The TAM information on the verb determines verbal inflection. Information specified in Dependency, Construction and Discourse level determines syntagmatic relation among the words. Finally Agreement rules adjust the final word forms as and when necessary.

R(ow)2	Concept	Śāma_1 / evening_1	eka_1 / one_1	do_1 / two_1	camaka_1 / shine_1	tārā_1 / star_1	najara+ā_1-tā_hai_1 / appear_1-pres
R3	index	1	2	3	4	5	6
R4	Sem Category	time					
R5	Morpho-	[- sg a]				[- pl a]	
R5	semantics						
R6	Dependency	6:k7t	5:card	5:card	5:rvks	6:k1	0:main
R7	Discourse						
R8	Speaker's view						
R9	Sentence type	affirmative					
R10	Construction	disjunct:[2,3]					

Table 2: Sent-1: USR for Sentence: 3a. In the USR -k7t = temporal, card = cardinal, rvks = relation vartamān kāl samānādhikarana-(present simultaneous time), k1 = kartā (close to agent but not completely equivalent)

R2	Concept	kucha_1	samaya_1	tārā_1	saṃkhyā_1	baḍha_1- tā_hai_1
R3	index	1	2	3	4	5
R4	Sem Category					
R5	Morpho-semantics		[- sg a]		[- sg a]	
R6	Dependency	2:quant	5:k7t	4:r6	5:k1	0:main
R7	Discourse			Sent-1.5:coref		Sent-1.6:contrast
R8	Speaker's view	1:emph				[shade:completion]
R9	Sentence type	affirmative				

Table 3: Sent-2: USR for Sentence: 3b. In the USR - quant:quantity, r6 = genitive, emph= emphasis,Light verb jā (go) adds a sense of completion to the main verb

3 Related Works on Evaluation

Evaluation of Semantic Representations is a multi-dimensional task due to many qualitative parameters that need to be evaluated. Usual parameters of interest are the utility of the semRep, invariance, Universality (cross-lingual potential), usability, computational efficiency etc (Abend and Rappoport, 2017).

Human evaluation is one of the important methods for measuring the accuracy of generation tasks. A human evaluator can determine the accuracy, give a qualitative ranking based on the naturalness/fluency as well as completeness of information encoded in a given semantic representation. Several human evaluation based methods are in practice such as the WMT tasks (Bojar et al., 2016), Direct Assessment (Graham et al., 2017), HUME (Birch et al., 2016) for UCCA, HTER (Snover et al., 2006), or SMATCH (Cai and Knight, 2013) applicable to AMRs.

Human evaluations, besides being more accurate for SemRep evaluations, can also mark strengths and weaknesses of the generation, further indicating possible improvements. However, human evaluation would usually require skilled annotators as well as well-designed annotation guidelines to ensure objectivity. Hence, human evaluation is effective but can be resource and time-inefficient (Sai et al., 2020). Human evaluation reliability and consistency are measured through Inter-Annotator Agreement (IAA). Automated evaluations are the alternative to human evaluations, as they can be consistent, as well as resource efficient. However, the notion of semantic similarity is still not fully captured by the machine. Several word based, vector based and embeddings based measures are available for the same (Sai et al., 2020).

In this paper, we attempt to strike a balance between both human and automatic evaluation of USR and propose two kinds of evaluation: (a) Qualitative and (b) Quantitative. Table 4 summarizes our evaluation.

4 Quantitative Evaluation

This paper presents the quantitative evaluation metrics of USR from two perspectives: (a) the Reliability of USRs; (b) the utility of USR in the domain of multi-lingual generation.

The reliability is evaluated through Inter-Annotator Agreement. The utility of USR is evaluated by examining the textual similarity between the reference sentence and the manually generated

Type	Exp Name	Quality parameter	Dataset	Measure
Quantitative	IAA	Reliability	Geo_simple	Human Evaluation - Agreement %, Cohen's kappa
Quantitative	NLG utility	Correctness, completeness	Geo_6	Pairwise cosine with embeddings
Qualitative	Generation experience	Usability/ Readability	Geo_6 + verified_sentences	Human evaluation - effort, difficulty level

Table 4: USR Evaluation Framework

sentence. Essentially this becomes an evaluation of the generation task (Abend and Rappoport, 2017). Further, the generation task can be used to examine the utility of USR for multi-lingual generation, This is an important quality to evaluate as USR is designed to facilitate Natural Language Generation in multiple languages by using the multi-lingual concept dictionary to find equivalent concepts and can generate the same thought in multiple target languages.

We have extensively used the idea of semantic textual similarity (STS) in our evaluations, measured through human evaluation as well as by standard measures like pairwise cosine similarity. Here, we build a USR for a reference sentence R, then use that USR to either manually or automatically generate a sentence (G). If R and G are semantically close, we can say that the USR correctly and adequately captures the reference sentence meaning. Table 4 summarizes our evaluation framework.

4.1 Measuring Reliability of USR

This section describes the Reliability i.e. Inter annotator Agreement experiment.

4.1.1 Dataset

Geo_simple is a corpus of 90 simple sentences (with a total word count 928) created from the Indian NCERT Geography textbook for grade 6 and grade 7. The average length of these sentences is 11 words. These sentences are simple sentences, with one finite verb and zero or more non-finite verbs. Complex sentences are manually simplified to create simple sentences with proper connectives.

4.1.2 Experiment Setup

An annotation guideline document (USR Guidelines) is provided to two expert annotators with more than 6 months of experience with USR and its annotation. Geo_simple_0 is a set of base USRs automatically generated from sentences in Geo_simple dataset. Annotators independently develop their own versions of the USRs by editing the USRs in Geo_simple_0. Inter-annotator agreement (IAA) for different semantic features (the rows of the USRs) is calculated and then aggregated for the three levels of semantic information captured in USR.

For certain type of sentences, the annotators can differ in the number of concepts they identify. One case is the annotation of complex predicates. A complex predicate is a Noun+Verb construction. There can be disagreement among the annotators on when to call a Noun followed by Verb construction a complex predicate and when verb-object construction. Depending on that decision, the number of concepts identified for a given USR changes among annotators such that the concepts and their indices may differ partially, resulting in two very different looking, but valid USRs. To handle these kinds of situations, IAA is calculated for two different cases: a) Match cases - the number of concepts match (b) Not match - the number of cases differ. About 25% of our Geography data exhibits a difference in the number of concepts identified for the same reference sentence.

Inter Annotator Agreement (IAA) is measured using Agreement Percentage as well as Cohen's Kappa for Match cases (Cohen, 1960), but only Agreement Percentage (Given as Partial Agreement) for Not Match cases as Cohen's Kappa will be appropriate for such cases. IAA is interpreted using the agreement schema given by Landis and Koch (Landis and Koch, 1977) for sentences. The result is given in the next section.

4.1.3 IAA Results and Discussion

We have calculated the Inter Annotator agreement (IAA) separately for 'Match cases' and 'Not match cases'. The 'match' and 'Not match' cases for both

Type	Match cases		Non match Cases
Feature category	**Cohen's Kappa**	**Agreement %**	**Partial Agreement %**
lexico-conceptual	0.898	92.13	73.74
Syntactico-semantic	0.758	92.50	43.85
Discourse	0.869	95.52	77.78
Sentence type	0.929	95.588	76.00

Table 5: A summary of agreements for Match and Non match cases.

data are given in Table 5 .

Maximum impact of 'Not Match' concepts is seen at the syntactico-semantic level mainly for dependency attachments (Table 5) due to change in index numbers of concepts, as number of concepts is different. For Match cases, the Cohen's kappa scores for gender and number are comparatively low (0.76, in Table 5). A detailed analysis shows that the disagreement in the lexico-conceptual category are mainly seen in the semantic category and GNP information. The GNP information shows disagreement mostly for pronominal concepts. It can be attributed to the lack of context. For example, in the following case, Annotator1 chose to consistently not mark the gender for pronominal terms while annotator2 has decided otherwise. See the following example: 2nd person pronoun *tuma* (you) can be both singular and plural in number. In such cases, annotators can overlook larger discourse information and tend to mark either singular(sg) or plural (pl) thus resulting in a disagreement in the annotation. Another low score in Table 5 is related to discourse relation. For this case, the agreement % is high while the Kappa score is comparatively low. Kappa is reducing the scores by assuming a probability of chance agreement, which itself has a low probability in our annotation exercise owing to the experience and expertise of our annotators. Hence, we feel that agreement % is a better measure of IAA for our annotations as compared to Cohen's Kappa. Results from the IAA experiment establish that the USR Guidelines is a reliable document and following that annotators with some training can reliably create USRs.

original_sentence	Annotator1	Annotator2
maiṃ bhī jāūṃgā	[- sg u],	[m sg u],

Table 6: GNP annotation differences in USR annotation

4.2 Measuring utility of USR for Multi-lingual generation

The utility of USR for multi-lingual generation is evaluated through a detailed experiment, where human generators manually parse the USRs to generate corresponding natural language sentences in Hindi, Bangla and Telugu by the aid of the multilingual concept dictionary. The underlying idea is as follows: If a generated sentence G (from USR U) and reference sentence R exhibit a high semantic textual similarity (STS), such that the USR U is created from R and is used to generate G, then it can be inferred that the semantic information captured by the USR is correct as well as adequate. The concept dictionary provides the corresponding concept in the desired output language. The generated sentences are evaluated manually and automatically for Semantic Textual Similarity.

4.2.1 Datasets

Geo_6 - The dataset consists of a corpus of 125 sentences from a Geography textbook of grade 6. These are simple sentences and do not contain any connectives. Complex sentences, if any are manually simplified to create simple sentences. The average length of these sentences is 11 words. Sentences from Geo_6 are used to programmatically generate a set of USRs (USR_0). The USRs are verified and edited by the experts for the correctness of content and structure (USR_1). USR_1 is used by a set of human generators to generate Hindi, Telugu and Bangla sentences.

Item	Score
Same meaning(Totally)	3
Minor difference in meaning	2
Not same at all	1

Table 7: Scoring Rubric for Human Evaluation of Semantic Textual Similarity

4.2.2 Multi-lingual generation Experiment Setup and Measures

All human generators, who are native speakers of their respective languages, are pre-trained to read USRs and decode the semantic information. The basic process for sentence generation in a target language is simple. For every reference sentence R the corresponding USR is made available to the human generator who manually parses the USR text structure. Human generators were asked to pay more attention to preserving information as it is (from USR) in the generated sentences and not to worry too much about maintaining the naturalness/fluency of the target language.

Once the human generators manually generate the sentences, a sanity checking is done in the following way before the automatic comparison with the reference sentences.

1. Reference Sentences without a corresponding generated sentence are excluded from further analysis.
2. Spelling mistakes are ignored.
3. Generated sentences with partially matching semantics are included in the response set, as they may indicate a deficiency in the USR.

For each sentence pair (Ri and Gi), we compute the Semantic textual Similarity (STS), manually as well by using known measures such as pairwise cosine measure after embedding sentences Ri and Gi using the state of art LaBSE model (Feng et al., 2020) as well as XLM-R (Conneau et al., 2020), a popular multilingual Masked Language Model (MLM). The embeddings are the vector representations of sentences such that the semantically similar sentences are closer, even if they belong to different languages, hence providing the cross-lingual measurement of similarity. The embeddings done using LaBSE provide reliable pairwise cosine measure (Feng et al., 2020).

Human evaluation of STS is done using the following scoring rubric (Table 7):

4.2.3 Results and Analysis

Hindi sentences are generated by two human generators. Hence we computed the internal consistency/reliability of human evaluation scores. The generations are internally consistent, and are acceptable as indicated by for human_generator1 (Cronback's Alpha score 0.76) and good for human_generator2 for (Cronback's Alpha score 0.82). Next, we compute the frequency distribution of STS scores from human evaluation across the three target languages Hindi, Bangla and Telugu (Table 8).

Next, we compute the pairwise cosine similarity, with embedding, for the four sentence pairs namely Ref-Hindi1, Ref-Hindi2, Ref-Telugu and Ref-Bangla. (Table 9) records our results of both human and automated evaluation.

As evident from the high scores given by the human evaluators (Table 8, Table 9), and by both the reasonable cosine similarity scores, (Table 9), we can conclude that the semantics are preserved in the USR by a high degree of accuracy. The scores are also reliable as we can see a similar pattern in the scores gained from the above three methods. Since the Semantic Textual Similarity is reasonably high across the three languages, we can also confirm the universal nature of USR.

The Inter-Annotator Agreement scores make it evident that USR is a reliable semantic representation. Similarly, utility of USR for multi-lingual generation is high due to the ease of rules-based parsing of USR to construct a meaningful sentence.

5 Qualitative Evaluation

It is important to understand and record the experience of people involved in creating and using USRs. We are particularly interested in the readability of USR, because the idea is to create a gold standard USR bank which is only possible when human annotators can effortlessly read USR and correct it as needed. In this paper, readability is tested in terms of correctness and ease of generating sentences from a USR. If a human generator succeeds in generating a correct sentence with minor or no assistance, that shows that the USR is readable as well as adequate for correct sentence generation. We conducted a study and the following survey to check the readability of USRs by human beings. Human generators (14) with mixed prior knowledge and experience with USRs are given the manual generation task. The experience distribution of generators is as given in Table 10

Each human generator was first trained on generating a sentence from a given USR. USR guidelines were explained to them and they practiced on 3 USRs. Then each generator was given a set of 10 USRs from Geo_6 and another dataset to independently generate Hindi sentences. They could refer to the USR guidelines as many times as required.

STS Score	Hindi		Bangla		Telugu		Total
Score	Count	% within Hindi	Count	% within Bangla	Count	% within Telugu	
3 (Totally)	633	84.40	70	76.92	25	60.98	728
2 (partially)	92	12.27	20	21.98	13	31.71	125
1 (not at all)	25	3.33	1	1.10	3	7.32	29
Count Sentences	750(2 sets)	100	91	100	41	100	882

Table 8: Frequency Distribution of Semantic Similarity scores (Human Evaluation)

	Ref-Hindi1	Ref-Hindi2	Ref-Telugu	Ref-Bangla
sentence	91	91	41	93
Human evaluation (average) - (0- 3 rating)	2.81	2.85	2.71	2.33 0.778/
Pairwise cosine with LaBSE embeddings (0-1.0)	0.884	0.9041	0.746	0.604
Pairwise cosine with XLM-R (0-1.0)	0.916	0.938	0.738	0.705

Table 9: Semantic closeness scores for Multi-lingual generation from USR

	Academic Degree in Linguistics or language		Any other Degree	
	Experience < 3 months	Experience > 3 months	Experience < 3 months	Experience > 3 months
Count of human generators	2	5	5	2

Table 10: Experience distribution of Human Generators

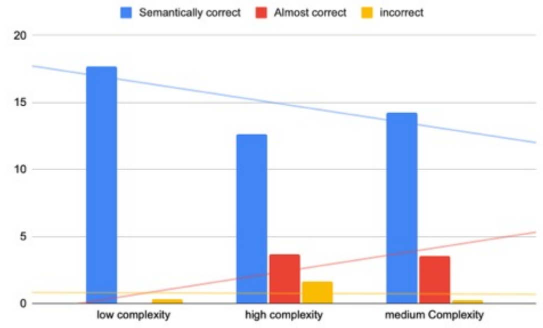

Figure 1: Generation correctness Vs. The complexity of the USR

The generators filled out a survey immediately after the Hindi generation task. The USRs were classified by the complexity level as low, medium and high, based on the number of concepts, and variations in dependencies, discourse, speaker's view information.

STS scores, measuring accuracy, for reference sentence and generated sentence were computed. A cross-sectional view of the correctness vs the complexity level is given in Figure 1. It is evident that generators could produce a high number of semantically correct (same meaning, and minor variations in meaning) sentences. The errors seen were mostly missing terms like 'almost, 'may-be', GNP and TAM (past vs present) variations. For example: For the reference sentence (Translated): *Sun is about 15 million KM away from the Earth*. Some generators did not include the word 'about'.

Figure 2 clearly indicates that the human generators could find the desired help in the USR guidelines. Most human generators found the USR Guidelines exhaustive and could use the document to clarify their doubts. The help was mostly sought for the dependency relations, as the list of dependencies is exhaustive, and remembering all can be an arduous task for a novice. Of the reported consultation of the USR guidelines, novice generators with < 3 months of exposure to USR required the most help as expected. The generators were also

		Difficulty Level					
		Very Easy	Easy	Ok	Difficult	Very Difficult	Total
My exposure to USR	<3 mth	0%	21.43&	14.29%	21.43%	0%	57.14%
	>3 mth	21.43%	7.14%	7.14%	7.14%	0%	42.86%
	Total	21.43%	28.57%	21.43%	28.57%	0%	100%

Table 11: Difficulty Level of USR Generation Process

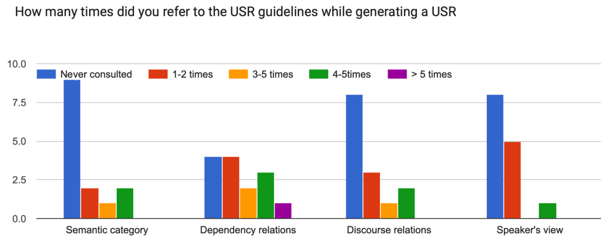

Figure 2: Frequency distribution of USR referrals while generating 10 USRs

asked to rate the difficulty of the generation process (Table 11). Majorly, the generators find the USR generation process to be very easy, easy, or OK (computed for both categories, <3 months exposure to USR; > 3 months exposure, using a Likert scale of 1-5, 5 being very difficult).

Based on the above experiences of the human generators, we can say with confidence that the readability of USRs is high as the generators could generate the USRs with ease, find the desired help in the guidelines, and could generate a high number of correct USRs. It is clear that the USR generation task is also not very difficult and gets easy with minor training. One important utility of USR readability measures is reflected in one of the tasks that we have taken up, namely training school children to read and write USR as an approach towards learning Universal Semantic Grammar through USR. The idea is that the USRs can enable children to overcome language barriers in communication.

6 Conclusion

In this paper, we have introduced a new SemRep called Universal Semantic Representation (USR). This is a very new initiative that attempts to capture the speaker's vivakṣā and is inspired from Indian Grammatical Tradition. The Lexico-Conceptual, Syntactico-Semantic and Discourse level information is encoded in a structured format in which USRs are interlinked to express the meaning of discourse as a whole. This paper presents the design of the USR and also records its detailed, multi-dimensional evaluation for reliability and its utility for natural language generation. Empirical evidence suggests high reliability as well as reliable semantic similarity scores for natural language generations done in multiple Indic languages namely Hindi, Bangla and Telugu. The qualitative evaluation strongly suggests that USR is easy to read and use with some training. Thus USRs are suitable for Natural Language Generation tasks, and can be used as a universal semantic representation.

7 Acknowledgement

We are thankful to every member of the Language Communicator team, specifically Arjun, Isma, Shweta, Bidisha and Hymavathi, for their contribution in data preparation and experiments. We are grateful to MEITY, Govt. Of India, for supporting and funding the Language Communicator project.

References

Omri Abend and Ari Rappoport. 2013. Ucca: A semantics-based grammatical annotation scheme. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013)–Long Papers*, pages 1–12.

Omri Abend and Ari Rappoport. 2017. The state of the art in semantic representation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 77–89.

KV Abhyankar. 1977. A dictionary of sanskrit grammar, (1: 1961). *Baroda.(= Gaekwad's Oriental Series 134).*

Collin F Baker, Charles J Fillmore, and John B Lowe. 1998. The berkeley framenet project. In *COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics*.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract meaning representation for sembanking. In *Proceedings of the 7th linguistic annotation workshop and interoperability with discourse*, pages 178–186.

Bhojaraja. 2007. *Shringaraprakasha*, volume 1. Motilal Banarsidass Publishers Pvt. Ltd. Delhi.

Alexandra Birch, Omri Abend, Ondřej Bojar, and Barry Haddow. 2016. HUME: Human UCCA-based evaluation of machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1264–1274, Austin, Texas. Association for Computational Linguistics.

Alena Böhmová, Jan Hajič, Eva Hajičová, and Barbora Hladká. 2003. The prague dependency treebank: A three-level annotation scenario. *Treebanks: building and using parsed corpora*, pages 103–127.

Ondrej Bojar, Christian Federmann, Barry Haddow, Philipp Koehn, Matt Post, and Lucia Specia. 2016. Ten years of wmt evaluation campaigns: Lessons learnt.

Shu Cai and Kevin Knight. 2013. Smatch: an evaluation metric for semantic feature structures. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 748–752, Sofia, Bulgaria. Association for Computational Linguistics.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale.

Ann Copestake, Dan Flickinger, Carl Pollard, and Ivan A Sag. 2005. Minimal recursion semantics: An introduction. *Research on language and computation*, 3:281–332.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic bert sentence embedding. *arXiv preprint arXiv:2007.01852*.

Yvette Graham, Qingsong Ma, Timothy Baldwin, Qun Liu, Carla Parra, and Carolina Scarton. 2017. Improving evaluation of document-level machine translation quality estimation. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 356–361, Valencia, Spain. Association for Computational Linguistics.

Paul R Kingsbury and Martha Palmer. 2002. From treebank to propbank. In *LREC*, pages 1989–1993.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

M OnlinePalmer, D Gildea, and P Kingsbury. 2005. The proposition bank: A corpus annotated with semantic roles. *Computational Linguistics Journal*, pages 31–1.

P Raster. 2015. *The Indian Grammatical Tradition*, volume 1. De Gruyter Mouton.

Ananya B. Sai, Akash Kumar Mohankumar, and Mitesh M. Khapra. 2020. A survey of evaluation metrics used for nlg systems.

Petr Sgall, Ján Horeckỳ, Alexandr Stich, and Jirí Hronek. 1992. Variation in language. *Variation in Language*, pages 1–381.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

Srisa Chandra Vasu et al. 1891. *The Ashtadhyayi of Panini*, volume 2. Satyajnan Chaterji.