UNIVERSAL SEMANTIC REPRESENTATION GUIDELINE VERSION 4.2

SENTENCE SEGMENTATION

USR: A brief outline

Universal Semantic Representation (USR) is a meaning representation that models Indian Grammatical Tradition (IGT). Meaning (or thought) is there in the mind of the speaker (author) and while speaking (writing), (s)he makes use of language (or linguistic expressions) to express his/her thought. Thus a discourse (text) represents the speaker's thought.

This guideline is created to help annotators to make USRs of the written discourse. The objective is to generate multiple natural languages from these USRs using Natural Language Generators.

Motivation of USR

Written text expresses the speaker’s intention of how (s)he wants to depict a situation. A situation can be seen as an event with various participants involved in that event and also various associated events either causing or resulting or associating with the main event. For example, let us take a situation where the main event is *speaking*. Two participants involved are *Ram* and *Sita* in the role of speakers. The location of the event is *bus stop*. The target person speaking is the *brother* of *Ram* and *Sita*. When the speaker wants to talk about this situation (s)he has to choose a tense and aspect. For example, the chosen time is past perfective in this case. This very basic situation (which we can call 'propositional information') can be expressed in Hindi as

(1) rāma aura sitā ne basa aḍḍe para bhāī ke sātha bāta kī.

Now, the speaker wants to add some more situational information on this basic propositional information. They are the following: the *brother* is younger to *Ram* and *Sita*. The speaker wants to negate the whole situation. In Hindi, the chosen linguistic elements are negation marker *nahīṃ* ‘not’. Moreover, the speaker wants to add the information of certainty to the negation of the above situation. However (s)he wants to leave open the possibility of the agents’ speaking to somebody else in the bus stop. Such intention of the speaker can be expressed through the discourse particle *to* in Hindi. Thus the exemplify sentence generated in different languages including Hindi is as below-

Language	**Expected Outcome of the Sentence**
Hindi	rāma aura sitā ne basa aḍḍe para apane choṭe bhāī ke sātha to nahīṃ bāta kī.
Bangla	rāma āra sitā bāsa sṭaiṃḍ-e nijera choṭa bhāīera sāthe to kathā bal-e ni.
Nepali	rāma ra sitā-le basa-bisaunī-mā āphno sāno bhāī-samga ta kurā gare-nan

Telugu	rāma sitā basa sṭaiṃḍ-lo vāīyīya cinna tammu-du-to ayite mātlāda ledu
Punjabi	rāma te sitā apane vīra nāla te basa sṭaiṇḍa to gala ni karyā
Marathi	rāma āñi sītene basasthāNakāvara apalyā choṭyā bhāvāSi tar nāhī bolale.
Tamil	rāma un Sita vum nichayama avunga thambi kitta pesavaeilla
English	Ram and Sita did not certainly talk to their younger brother at the bus stop.

Table 1. Example of expected generated sentences in different languages from a given USR

There can be one more interesting interplay of negation and certainty information in this case. The speaker here wants to say that (s)he is certain that Ram and Sita did not talk to their younger brother in the bus stop. Thus certainty takes a wider scope on negation of the actual event of Ram and Sita's speaking with their younger brother at the bus stop. Instead, if the speaker wanted to express that he is not certain if Ram and Sita spoke to their younger brother in the bus stop, then the semantics of negation *nahīṃ* would take the wider scope over *to* 'expressing certainty'.

In both cases, the sentence generated would have been the same. However, in USR, we have the opportunity to specify the scopal information. The speaker can annotate the appropriate scopal order of negation and discourse particles to express what (s)he actually means.

A text contains a series of sentences. Sometimes, the relation among the sentences are explicitly marked through discourse markers. These discourse markers maintain the flow of the story. For example, the speaker in this case might want to justify why (s)he assumes that *Ram* and *Sita* did not speak with their younger brother that day. In order to express that thought, the sentence generated can be:

(2) Hindi: kyoṃki usa dina unakā bhāī śahara meṃ thā hī   nahīṃ
Bangla: kāran sedina   oder   bhāi sahar-e   chi-lo-i nā

kyoṃki 'because' is a discourse connective marker that logically connects (1) and (2) by justifying (1) through (2). *usa*- and *una*- (pl of *usa*-) are anaphoric pronouns. *usa dina* refers to the same day when the event took place. *una* in *unake bhāī* refers to *Ram* and *Sita*. These anaphoric expressions are the mechanism for maintaining the cohesiveness in the story. The discourse particle *hī* again like *to* in (1) add extra-propositional meaning which actually conveys the speaker's view or perspective.

USR attempts to capture all this information in a human-friendly yet machine tractable representation.

Sentence Segmentation

Since USR annotation of complex sentences is difficult and automated USR generation for complex sentences is a challenge as observed through several experiments, we have decided to first segment complex sentences into discourse units without losing information. Some complex sentences are not segmented as segmenting them will make the discourse less coherent.

Following are the strategies of sentence segmentation

- In general, segmented segments will be a discourse unit which contains a finite verb.
- A discourse unit is a simple sentence or a clause which is not necessarily the smallest unit. It participates in making the larger discourse.

Such as- rāma aura sitā ne basa aḍḍe para bhāī ke sātha bāta kī. ‘Ram and Sita spoke to their brother in the bus-stand.’

- Relative Clauses with the relative pronoun referring to a noun in the sentence are not segmented. Such as -

bhārata kā sabase dakṣiṇī biṃdu jo iṃdirā biṃdu kahā jātā thā, san 2004 meṃ jalamagna ho gayā.

‘The southernmost point of India, which was known as Indira point, was submerged in water in the year 2004.’

This sentence is not split.

When to split Relative Clauses:

1. If a sentence contains more than one relative clause, relative clauses are segmented and their inter-clausal relations are shown in discourse element row. Such as-

Sent_ID_1	pṛthvī ke dharātala ke ūṁce uṭhe hue bhāga jinakā śikhara hajāra mīṭara se adhika ūṁcā ho aura ḍhāla tīvra ho, tathā jinake banane jinakā lākhoṃ varṣa lage, parvata kahalāte haiṃ|

The above sentence contains more than one relative clauses and they will be segmented as following

-

Sent_ID_1a	pṛthvī ke dharātala ke ūṁce uṭhe hue bhāga parvata kahalāte haiṃ
Sent_ID_1b	jinakā śikhara hajāra mīṭara se adhika ūṁcā ho
Sent_ID_1c	aura jinakā ḍhāla tīvra ho
Sent_ID_1d	tathā jinakā banane me lākhoṃ varṣa lage

See Relative Clause for annotation rules

2. If a relative pronoun functions as a discourse connective, those relative clauses will be splitted. Such as,

nadī ke nicale bhāgoṃ meṃ ḍhāla kama hone ke kāraṇa nadī kī gati kama ho jātī hai, jisake pariṇāmasvarūpa nadīya dvīpoṃ kā nirmāṇa hotā hai.

Here, the whole expression jisake pariṇāmasvarūpa acts as a discourse connective. Hence, the clause it is attached with, is splitted from the previous clause it is connecting with and the two sentences will be:

nadī ke nicale bhāgoṃ meṃ ḍhāla kama hone ke kāraṇa nadī kī gati kama ho jātī hai.
isake pariṇāmasvarūpa nadīya dvīpoṃ kā nirmāṇa hotā hai.

Strategy for splitting complex sentences:

- Complement clauses will be splitted following the rules stated below-
  A. sentential or clausal complement will be an independent sentence.
  B. yaha ‘this’ will be added with the clause containing the main verb.
  C. yaha ‘this’ will co-refer the entire complement clause.see here for detail.

Original Sentence

Sent_ID_1	# hīrā ne kahā ki ūṁṭa mileṃge. ‘Hira said that the camel will be available there.’

After segmentation

Sent_ID_1a	hīrā ne yaha kahā ‘Hira said this.’
Sent_ID_1b	ūṁṭa mileṃge ‘Camel will be available there.’

Complement Clause may occur as following -

Original Sentence

Sent_ID_1	# hīrā ne **itanā** kahā ki ūṁṭa mileṃge. ‘Hira said that the camel will be available there.’

We adopt the strategy of segmenting such sentences as following

Sent_ID_1a	hīrā ne **itanā** kahā ‘Hira said this.’
Sent_ID_1b	ūṁṭa mileṃge ‘Camel will be available there.’

- **itanā…ki as discourse connective**

itanā…ki may occur as a discourse connective as well. We segment them as following -

Original Sentence

Sent_ID_1	#nadī ke bāhya taṭa yā natodara taṭa kā itanī tejī se aparadana hotā hai ki visarpa lagabhaga pūrṇa vatta bana jātā hai|

We split such sentences and postulate 'isase' as discourse connective in the segmented sentence which brings 'pariNama' relation and add iwanA_ki in the speaker's view row. See here for detailed USR annotation strategy.

After sentence segmentation

Sent_ID_1a	#nadī ke bāhya taṭa yā natodara taṭa kā tejī se aparadana hotā hai
Sent_ID_1b	#isase visarpa lagabhaga pūrṇa vatta bana jātā hai|

- When two clauses are connected with a connective, we split the sentence into two independent sentences and retain the connective in the sentence where it originally is.

Original Sentence

Sent_ID_1	# merī sāikila suṃdara hai lekina abhī vaha gaṃdī hai 'My cycle is beautiful but it is dirty now.'

After sentence segmentation

Sent_ID_1a	# merī sāikila suṃdara hai 'My cycle is beautiful'
Sent_ID_1b	# lekina abhī vaha gaṃdī hai. 'But it is dirty now'

Original Sentence

Sent_ID_2	#rām bīmāra hai isalie vaha skūla nahīṃ gayā 'Ram is sick. Therefore he did not go to school.'

After sentence segmentation

Sent_ID_2a	#rām bīmāra hai 'Ram is sick'
Sent_ID_2b	#isalie vaha skūla nahīṃ gayā 'He did not go to the school'

Original Sentence

Sent_ID_3	#rāma skūla nahīṃ gayā kyoṃki vaha bīmāra hai 'Ram did not go to the school because he is sick.

After sentence segmentation

Sent_ID_3a	#rām skūla nahīṃ gayā 'Ram did not go to school.'
Sent_ID_3b	#kyoṃki vaha bīmāra hai 'Because he is sick.'

- When two clauses are connected with a paired connective, we split the sentence into two independent sentences and retain the connective in the main clause.

Original Sentence

Sent_ID_4	# yadi āpa mujhe āmaṃtrita karate haiṃ to maiṃ āpake ghara āūṃgā 'If you invite me then I will come to your house.'

After sentence segmentation

Sent_ID_4a	#āpa mujhe āmaṃtrita karate haiṃ 'You invite me.'
Sent_ID_4b	#to maiṃ āpake ghara āūṃgā 'Then I will come to your house'

The annotation of discourse connective is presented in the Discourse Connective Relation section to ensure no loss of information.

USR_GUIDELINES - V 4.2