Cross-linguistic Semantic Tagset for Case Relationships

Ritesh Kumar[1], Bornini Lahiri[2], Atul kr. Ojha[3]
[1]Dr. Bhimrao Ambedkar University, Agra, [2]Central Institute of Hindi, Hyderabad
[3]Panlingua Language Processing LLP, New Delhi
[1]ritesh78_llh@jnu.ac.in, [2]lahiri.bornini@gmail.com,
[3]shashwatup9k@gmail.com

Abstract

In this paper, we discuss the development of a semantic tagset for annotating and disambiguating case relations across languages. The case relations are currently defined using the formal properties of case affixes or postpositions. However cross-linguistically (even in languages within the same language family or same sub-family), the semantics of case-bearing do not exactly overlap. While case-bearing units across languages may share a core semantics, the peripheral semantics differ to a large extent. This results in ambiguities in several multilingual tasks like machine translation, cross-lingual information retrieval and others. We propose a new tagset for semantic annotation of case-bearing units based on an inventory of the semantics that they express. We are currently experimenting with this tagset and machine translation tasks and the preliminary results show that such annotation could be useful in handling at least some of the ambiguous instances.

1 Introduction

Case and case relations are one of the most extensively studied, investigated and debated upon phenomena in almost every theoretical framework of Linguistics, which includes descriptive and typological framework (some of these include Schlesinger (1979), Nikiforidou (1991), Bickle and Yadava (2000), Malchukov (2008), Corbett and Noonan (2008), Ziga and Kittil (2010), Jussi et al. (2011), Aikhenvald (2013), Lahiri (2013), Sinnemki (2014), besides numerous others). While there is some general understanding about what case relationships are, there are as many different ways of understanding case as there are theoretical frameworks. On the other hand, within the broader field of Natural Language Processing, and especially the applications which involve more than one language, misinterpretation and ambiguities of case relationships are often a source of multiple errors (see Ojha (2018), Sukhda (2017), Avramidis and Koehn (2008) and others for a discussion on the issues related to the case relationships in machine translation systems). There have been different approaches that have been used for handling these errors within different approaches of machine translation. Sukhda (2017) discusses the issues in handling case relations within rule-based machine translation approach. Yamada and Knight (2001) proposes an automatic method to handle case-related issues in machine translation using the tree structure. (Vanmassenhove and Way, 2018) makes use of WordNet synsets to annotate the lexical items with their senses in order to handle different kinds of ambiguities.

However despite large number of studies in both the fields, there has been hardly any attempt to integrate the insights gained from the extensive typological studies on case for handling the errors related to case in bi/multilingual systems. In this paper, taking cues from Lahiri (2014), we develop an extensive tagset for annotating nouns, verbs as well as postpositions with semantic information that could help in describing and disambiguating the semantics of such units across languages. Lahiri (2014) discusses postpositions / case markers / relators [1] as essentially polysemous entities which is composed of different extensions of a core meaning across different languages, leading to the use of different relators in one language

[1] There is an intensive debate among linguists regarding the status of postpositions as case markers. While some linguists argue that case markers are only inflectional, others argue that postpositions perform the same function as the affixes for case. In order to avoid this debate, neutral terms like ’flags’ and ’relators’ have also been proposed. We will use one of these terms - relators - in this paper to refer to both the postpositions and case markers

for one single relator in another language. The study was based on 7 different Eastern Indo-Aryan languages and it was observed that even within this small group of languages, the semantics of relators is not uniform. The study uses cognitive framework to discuss the polysemy of relators across these languages and comes up with comparative semantic maps to represent the areas of overlap and departure across languages for the relators which are canonically assumed to be marking the same relationship. The study makes use of several different kinds of meta-terminology to describe and explain polysemy in the semantics of relators. The tagset discussed in the paper is derived from this and is expected to disambiguate the different kinds of semantics expressed by a relator in a particular language and map that to the relator expressing the same semantics in the other language.

2 The Tagset

Blake (2001) defines case as ”marking dependent nouns for the type of relationship they bear to their heads” (Blake, 2001, p. 1). Thus it is evident that the semantics of case / relators depend on those of noun and its head, which may be a verb or another noun (as in the case of genitives). Lahiri (2014) also shows that the semantics of relators depend on the type of verb and also the semantics of nouns they are occurring with. As such the semantic classification of verbs and nouns could also prove to be significant for teasing apart different semantic values of the polysemous relators.

The proposed semantic tagset for case relationships, thus, include semantic tags for nouns as well as verbs. These tags are not exhaustive and are not meant for marking general semantics of nouns and verbs. Rather they are devised for their relevance in disambiguating case relationships.

3 Application

We are currently using this tagset to annotate data for two different machine translation systems - English - Bhojpuri and English - Magahi. We plan to use the annotated dataset for training a new machine translation system and we hope to get improvement over the existing systems that are built without using these semantic information.

Semantic Relation	**Label**	**Noun**	**Label**
Agentive	AGN	Agent	AGE
EM [1]	EXP	Patient	PAT
GST [2]	L_GST	Experiencer	EXP
Top	L_TOP	Beneficiary	BEN
Under	L_UND	Instrument	INS
Beside	L_BES	Location	LOC
Behind	L_BEH	Goal	GOA
Outside	L_OUT	Source	SOU
Inside	L_INS	Stimulus	STI
Above	L_ABV		
Peripheral	L_PER		
Core	L_COR		
In between	L_BET	**Verb**	**Label**
Benefactive	BEN	Static	STA
Causative	CAU	Dynamic	DYN
Objective	OBJ		
Instrumental	I_INS		
Perlative	I_PER		
Ablative	A_ABL		
Comarative	A_COM		
Possession	G_POS		
Partitative	G_PAR		
Kinship	G_KIN		
Social term	G_SOT		
Ownership	G_OWN		
Worth	G_WOR		
Nominalizer	G_NOM		
Material	G_MAT		

Table 1: The Semantic Tagset for Case Relationships / Relators [[1]Experiencer Marker; [2]General Spatial Term]

References

Alexandra Y. Aikhenvald. 2013. Possession and ownership: a cross linguistic perspective. In Alexandra Y Aikhenvald and R.M.W. Dixon, editors, *Possession and Ownership: a cross-linguistic typology*, Explorations in Linguistic Typology (1-64). Oxford University Press.

Eleftherios Avramidis and Philipp Koehn. 2008. Enriching morphologically poor languages for statistical machine translation. *Proceedings of ACL-08: HLT*, pages 763–770.

Balthasar Bickle and Y.P. Yadava. 2000. A fresh look at grammatical relations in Indo-Aryan. *Lingua*, 110:343 – 373.

Barry J. Blake. 2001. *Case*, 2 edition. Cambridge Textbooks in Linguistics. Cambridge University Press.

Greville G. Corbett and Michael Noonan. 2008. *Case* *and Grammatical Relations: Studies in honor of Bernard Comrie*. John Benjamins Publishing Company, Amsterdam.

Ylikoski Jussi, Vsti Katja, and Kittil Seppo. 2011. *Case, Animacy and Semantic Roles*. John Benjamins Publishing Company, Amsterdam.

Bornini Lahiri. 2013. A typological study of local cases of EIA languages. *Journal of Universal Language*, 14(1):85 – 115.

Bornini Lahiri. 2014. *A Typological Study of Cases in Eastern Indo-Aryan Languages*. Ph.D. thesis, Jawaharlal Nehru University, New Delhi.

Andrej L. Malchukov. 2008. Animacy and asymmetries in differential case marking. *Lingua*, 118:203 – 221.

Kiki Nikiforidou. 1991. The meanings of genitive: A case study in semantic structure and semantic change. *Cognitive Linguistics*, 2:149 – 205.

Atul Kr. Ojha. 2018. *English-Bhojpuri SMT System: Insights from the Karaka Model*. Ph.D. thesis, Ph D thesis, Jawaharlal Nehru University, New Delhi, India.

I. M. Schlesinger. 1979. Cognitive structures and deep structures: The case of the instrumental. *Journal of Linguistics*, 15:307 – 324.

Kaius Sinnemki. 2014. A typological perspective on differential object marking. *Linguistics*, 52(2):281 – 313.

Sukhda. 2017. *A Paninian Perspective to Information Dynamics in Language: Mapping Structures between English and Hindi*. Ph.D. thesis, IIIT Hyderabad.

Eva Vanmassenhove and Andy Way. 2018. Supernmt: Neural machine translation with semantic supersenses and syntactic supertags. In *Proceedings of ACL 2018, Student Research Workshop*, pages 67–73.

Kenji Yamada and Kevin Knight. 2001. A syntax-based statistical translation model. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*.

F. Ziga and Seppo. Kittil. 2010. *Benefactives and Malefactives. Typological perspectives and case studies*. John Benjamins Publishing Company, Amsterdam.