ODIA CASE-MARKER AGREEMENT TO MULTI WORD EXPRESSIONS (MWE) FOR ENGLISH-ODIA MACHINE TRANSLATION SYSTEM

Rudranarayan Mohapatra

ABSTRACT

In the modern approach of shallow parsing technique, we continuously get to see instances of systems built with a mix of several techniques yielding better results for most of such engines. The term "multiword expression" denotes a wide range of linguistic constructions such as idioms, fixed phrases, noun compounds, compound and complex verbs etc. Irrespective of easily controlled by native speakers, their interpretation poses a major challenge for computational systems, due to their flexible and heterogeneous nature. The semantics of a MWE cannot be expressed after combining the semantics of its constituents. Therefore, here is the formalism of hybrid semantic clustering followed by knowledge-based approach is discussed as effective instrument for extracting MWEs especially for resource constraint and agglutinative languages like Odia.

Key Words: Multi Word Expression (MWE), Machine Translation (MT), VP-chain, Tense-Aspect-Modality (TAM)

1. Introduction

From the year 2014 Odia became the sixth language of the country to get the "classical language" status in the same league as Sanskrit, Tamil, Telugu, Kannada and Malayalam. The Chief Minister of Odisha interviewed to 'The Hindu' said, "It is matter of pride for Odisha across the globe. The classical language status will create more opportunities for research and development of Odia". The language belongs to the Indo-Aryan language of the Indo-European language family. It is an official Indian language, and it is spoken by approximately 31 million people (80% of total population of Odisha) in Odisha, and other neighbor states like West Bengal, Jharkhand, and Chhattisgarh and Andhra Pradesh. However the legendary language having a large vocabulary and lexical resources got its statuesque in Official form yet to struggle for its existence by deficiency of gradually decreasing the speakers. Multi Word Expression (MWC) is a chunk with two or more words, where each word links to a semantic head through different dependency relations and it is a bidirectional machine learning process; i.e. the chunk MWE

can be prepared in Source language in a machine learning process looking to the behavioral nature of target language and vice versa.

2. Literature Survey:

In Odia language a word may appear in more than one grammatical category and in a grammatical category a word can have multiple senses. All words which depict the same sense (same meaning) are grouped together to form a single entry in the Target language when Automatically translated from its source language like English, we treated them as Multi Word Expressions having unique sense or meaning. In the early 1990s, MWEs started receiving increasing attention in corpus-based computational linguistics and NLP in various languages like English, German and many other European languages. An unsupervised graph-based algorithm to detect compositionality of MWEs was also proposed in the research of Korkontzelos and Manandhar (2009, p. 65). Collins, Koehn and Kučerová in German-to-English SMT use a syntactic parser to obtain an analysis of the source language string, then they apply a series of transformations to the parse tree, effectively, reordering the source string. The goal of this step is to recover the word order that is closer to the target language word order than the original string.[3] Previously in the CoNLL conferences several different chunking-related tasks, such as text chunking, Phrasal identification [4] and semantic role labeling, were carried out.

3. Case study & Observations:

Oriya nouns have both singular and plural forms: 'pila (child)', pila-maane (children). Oriya has a rich case system, marking nominals for accusative/dative (ku), instrumental (re), ablative (ru), genitive (ra/nka), and locative (re/ri) cases. Nouns in the nominative are not marked. Case markers may be preceded by plural markers, or by the definite marker. With the base case marker, it represents to singular. When the direct object is specific, in that case the accusative case is used.

1. The Nouns at the time of Phrasal agreement takes singular in form but plural in sense. Five children › ପାଞ୍ଚଜଣ ପିଲା [*paanchajana pilaa],* (Here 'ପିଲା' *[pilaa]* (child) singular in form takes plural in sense looking to the Quantifier agreements). Again in Odia language replicating words having hyphened in between are definite MWEs. However the expressions those without hyphen may not be so. In the said case, the MWE expressions depends upon the contextual categories so sallow parsing in nature.

For an illustration in Odia sentence {Verb_(i-matra/u-matra)} => { of + ‹verb_x›+ing } in English Sentences. Where (i-matra/u-matra) are categorically denotes the marker of present participle/past participle. In the sentence *'Se daudu daudu padigalaa'*, the interpretation yields as 'He fell down of running'. Looking to the Odia language agglutination and Verb Phrase (VP) chaining we are trying to limit the discussion of VP phrase chunking in details to its next level.

2. There is a separate module should be used for Named entity recognition (NER). In the case of {NP + of + PNP} where PNP is belongs to NER and NP denotes Common NP of English NP phrase it either reduce the physical existence of 'of' in Odia language and chunks as {PNP + NP} after the swapping. The phrase 'Government of India' not as 'Bharatara sarakara' but as 'bharata sarakara'.
3. Idiomatic Compound Nouns having non-productive in nature (ex: [*pila-pili*) takes inflection only to the last word of MWE. The phrases like quantitative and numerical in nature are highly productive, impenetrable and allow slight syntactic variations like inflections. Inflection can be added only to the last component Ex: [*sadhe chha ghantal* ('six hours and thirty minutes').
4. Relational Compounds: The Noun phrases of this category are mainly kin terms, Ex: [*maamu pua*l (Uncle's son) and are highly productive in nature. Here the Inflection can be added to the last word of the MWE.
5. Reduplicated Terms and Mimics: Reduplications and Mimics are non-productive and tagged as noun phrases, ex: khat-khat, chid-chid etc.

4. Proposed Methodology

Generally Odia language is free phrase order in nature rather than the word order. So head driven shallow parsing with heuristically machine learning method would help to understand the MWE of this language to use it in Machine translation system English-Odia. So MWE can be considered in three different ways. First is the MWE from source language nature and convention, second one is from target language convention and grammatical in nature. Third one is most important and bidirectional in nature in machine learning process looking to transfer contextual and most effective sense and meaning of the text. The proposed system will use the three tire filtration module in order to reach a near-accuracy MWE and trying to limit in between the use of Tense-Aspect-Modality (TAM) features used in template structure. A template is a serialization of variables which represent the morphological form if any. The case markers are very much related to Noun phrase and irrespective of word order in source text the case marker in target text is well dependent on target language MWE. In English from sentence

"The river Ganga has five tributaries" the sentence, phrase 'The river Ganga' (The + river + Ganga) when localized to Odia language it would be 'ଗଙ୍ଗା ନଦୀର ପାଞ୍ଚୋଟି ଉପନଦୀ ରହିଛି ।' (*Ganga nadira paanchoti upanadi rahichhi).* Here the Odia phrase parallel to English Phrase 'ଗଙ୍ଗା ନଦୀର' *(Ganga nadira)* is (Ganga + river) after swapping. The proposed system architecture would use the Bayesian networks (Pearl 1985) to express multiple interdependent linguistically motivated features and they may pass through N-Gram model for combined feature identification for its case-marker agreement in target language. So for both the languages in hybrid Machine learning process will try to capture the order of tokens became change irrespective of their word to word phrasing. And the nominal case marker 'ra' takes the position right to 'river' not to 'Ganga' as source language. Based upon the training to heuristic approach, here the MWE of Odia language is very much interdependent to nature of source language irrespective of conventional monolingual MWE.

5. Conclusion:

In this paper, differed from conventional MWE identification and the agreement of case-markers, the template based inter-dependable MWE expression identification and looking its TAM features how the target language case-markers agreement varies in agglutinative language like Odia are discussed. The template based hybrid stratified approach to find out Multi-word Expressions with its intermingled modified features for its further processing of case-marker agreement.

Reference:

1. Shallow morphology based complex predicates extraction in Oriya, R.C. Balabantaray and et. al., IIIT BHUBANESWAR, International Journal of Computer Applications (0975 – 8887), Volume 16– No.1, February 2011
2. Bharati, Chaitanya and Singhal "Natural language processing a paninian perspective", PHI.
3. Using TectoMT as a Preprocessing Tool for Phrase-Based Statistical Machine Translation, Daniel Zeman, Univerzita Karlova v Praze, ÚFAL, Malostranské náměstí 25, 11800 Praha, Czechia, http://ufal.mff.cuni.cz/~zeman/
4. Erik F. Tjong Kim Sang and S. Buchholz 2000, Introduction to CoNLL-200 Shared Task: Chunking. In Proc. of CoNLL-2000 and LLL-2000. Lisbon.p127-132.
5. VP-Chaining in Oriya, Dorothee A. Beermann and Lars Hellan, NTNU, Trondheim, Norway, Proceedings of the LFG02 Conference, National Technical University of

Athens, Athens Miriam Butt and Tracy Holloway King (Editors) 2002, CSLI Publications, http:/csli-publications.stanford.edu/