INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY
HYDERABAD
R&D SHOWCASE 2021
Technology, Social Impact

Semantic Role Labelling on Hindi-English Code-Mixed Data

ABSTRACT

•A 'semantic role' is the underlying relation that a constituent has with the predicate in a sentence.
•We present a 2-step system for automated Semantic Role Labelling of Hindi-English code-mixed tweets.
•We explore the issues posed by noisy, user generated code-mixed social media data.
•We also compare the individual effect of various linguistic features used in our system.
• Our proposed model is a 2-step system for automated labelling which gives an overall accuracy of **84%** for Argument Classification, marking a **10%** increase over the existing rule-based baseline model.

Label	Description
ARGA	Causer
ARG0	Agent
ARG1	Theme
ARG2	Benificiary
ARG2_ATTR	Attribute
ARG2_LOC	Physical Location
ARG2_GOL	Destination
ARG2_SOU	Source
ARG3	Instrument
ARGM_DIR	Direction
ARGM_LOC	Location
ARGM_MNR	Manner
ARGM_EXT	Extent
ARGM_TMP	Temporal
ARGM_REC	Reciprocal
ARGM_PRP	Purpose
ARGM_CAU	Cause
ARGM_DIS	Discourse
ARGM_ADV	Adverb
ARGM_NEG	Negative
ARGM_PRX	Complex Predicate

Table 1:PropBank Tagset used for snnotation

APPROACH

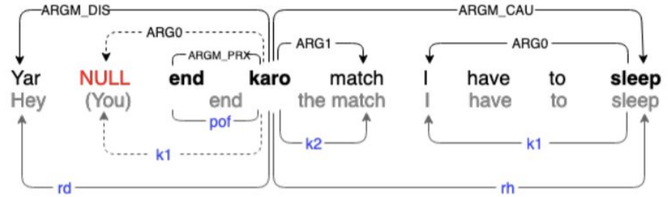

Figure 1:Example sentence from the corpus with Dependency and PropBank labels

The first step is to identify the arguments of the predicates in the sentence.These identified arguments are then classified into semantic roles in step 2.
We used SVM for the first step of binary classification
One vs rest multi class SVM was used for the second step. The data was split 80:20 for training and testing respectively.

Features Used

We used 14 linguistic features in our system.

Baseline features:
- Identified verb in the sentence
- Headword of the chunk
- Part of Speech tag of the head word
- Syntactic category of the phrase (NP, VP, CCP etc.)
- Predicate + Phrasetype
- Predicate + Headword

Indian Languages:
- Paninian dependency label
- Named Entities
- HeadwordPOS + Phrasetype
- Headword + Phrasetype
- HeadwordPOS(UD)
- UD dependency label

Code-Mixed data:
- Predicate + language
- Headword + language

RESULTS

Feature	Identification		
	P	R	f-score
Baseline	56	53	55
with predicate-lang	57	54	55
+dependency	81	76	78

Table 3:Accuracy scores for Argument Identification.

Feature	Classification		
	P	R	f-score
Baseline	27	15	19
+dependency	84	84	84

Table 4:Accuracy scores for Argument Classification.

The previously proposed rule based baseline model gives an accuracy of 96% and 73% respectively for Argument Identification and Classification respectively.
As the classification step is based on the identified arguments from the first step, we choose to adopt a hybrid approach. We use the rule based system for argument identification and SVM for argument classification

REFERENCE & CONTACT

- https://github.com/riyapal/Hi-En-SRL
- http://web2py.iiit.ac.in/research_centres/publications/view_publication/mastersthesis/833
- riya.pal@gmail.com

Authors: Riya Pal and Dipti Misra Sharma

Research Center Name: LTRC