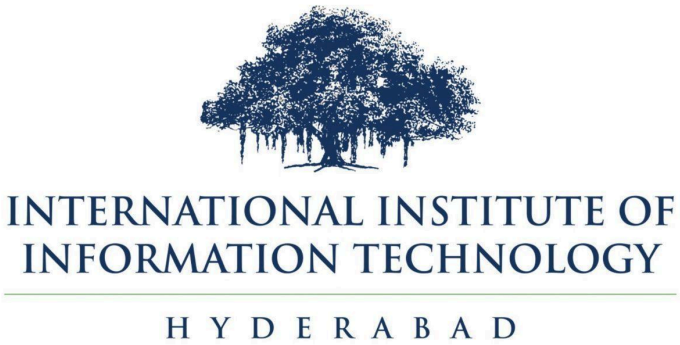
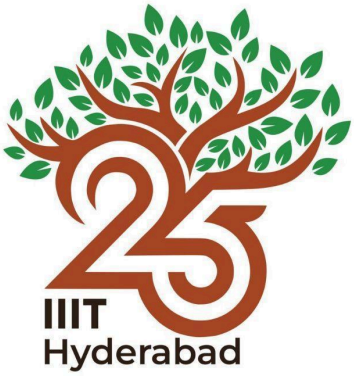




SEMANTIC ROLE LABELING
Final Submission - Project

COURSE: *INTRODUCTION TO NLP* - S24CS7.401

Advisor:
Prof. Manish Shrivastava and Prof. Rahul Mishra

Mentor:
Advaith Malladi

Team Number:
54

Team Name:
Lang3.1

Team Members:
Mohit Sharma - 2022201060
Neeraj Asdev- 2022201056
Hrishikesh Deshpande - 2022201065

Academic year:
2023-2024

1. Project Description :

Semantic Role Labeling (SRL) is a crucial task in Natural Language Processing (NLP) focused on identifying relationships between words or phrases in a sentence and their roles in events or actions. SRL assigns labels to these elements to explain who or what performs an action, what is affected by it, and other relevant details about the participants involved. This process is essential for a broad range of downstream NLP applications, including question-answering, inference, knowledge graph creation, and participant detection.

SRL approaches vary, including rule-based, supervised, and unsupervised learning. Rule-based methods use predefined rules and linguistic knowledge, while supervised methods use annotated data to train machine learning models. Unsupervised methods aim to identify semantic roles without explicit annotations. The Semantic Role Labeling (SRL) project aims to accurately identify and assign semantic roles to words or phrases in sentences and improve the performance of existing SRL systems. The project addresses the challenge of precisely capturing word-role relationships across diverse linguistic contexts, particularly in the Hindi language.

The project aims to develop and implement a Semantic Role Labeling (SRL) system for Hindi or Indian languages, using one or more available approaches, including rule-based, supervised, or unsupervised methods. Various factors, such as the availability of annotated data, the complexity of the Hindi language, and emerging challenges specific to Hindi, may influence the project's scope. A primary objective is to build a model, whether statistical or neural, capable of automatically labeling the arguments for each predicate in a Hindi sentence. This requires analyzing Hindi sentences to identify the roles of individual participants, using a combination of Hindi part-of-speech tags, dependency labels, and other linguistic information essential for determining theta roles and associated details.

2. Datasets:

Hindi Propbank: This dataset, specifically designed for Semantic Role Labeling tasks in Hindi, is a part of the Hindi-Urdu PropBank project. The PropBank aims to create a large-scale resource for training machine learning algorithms and conducting corpus linguistics studies in Hindi-Urdu. It involves building a multi-representational and multi-layered Treebank for Hindi-Urdu, addressing the scarcity of computational resources for the language.

We used it to prepare our dataset for training and testing our models. We also created a dataset from here , comprising around 14,000 tokens of Hindi text along with additional information like head POS , dependency. Additionally, we collected a set of

1.3k Hindi sentences along with arguments, SRL labels. Both of these datasets were utilized in training and testing our models.

Some of the semantic role label tags, dependency labels and their meanings :

Label	Description
ARG0	Agent, Experiencer or doer
ARG1	Patient or Theme
ARG2	Beneficiary
ARG3	Instrument
ARG2-ATR	Attribute or Quality
ARG2-LOC	Physical Location
ARG2-GOL	Goal
ARG2-SOU	Source
ARGM-PRX	noun-verb construction
ARGM-ADV	Adverb
ARGM-DIR	Direction
ARGM-EXT	Extent or Comparision
ARGM-MNR	Manner
ARGM-PRP	Purpose
ARGM-DIS	Discourse
ARGM-LOC	Abstract Location
ARGM-MNS	Means
ARGM-NEG	Negation
ARGM-TMP	Time
ARGM-CAU	Cause or Reason

Dependency Labels	Description
k1	karta - doer/agent/subject
k2	karma - object/patient
k1s	noun complement
k4	sampradana - recipient
k3	karana - instrument
k4a	experiencer
k2p	goal, destination
k5	apadana - source
k7	location elsewhere
k7p	location in space
k7t	time
adv	adverbs
rh	reason
rd	direction
rt	purpose

3. Challenges: One of the significant challenges we encountered was the scarcity of suitable datasets for training our models, particularly for Semantic Role Labeling (SRL) tasks in Hindi. Despite extensive search efforts, we found a lack of readily available datasets specifically tailored for this purpose. Additionally, the existing work on SRL in Hindi was limited, and we encountered a shortage of resources and references in this domain.

4. Experiments:

1. **Dataset Preparation :** To prepare our dataset, we began by collecting data from sources such as Propbank, available in a tree-like structure from GitHub of previous similar works. From this dataset, we filtered out relevant tokens for our experiments, ensuring to gather additional information about each token.

   Additionally, we extracted sentences from the Propbank dataset, making sure to retain all associated labels, including argument POS tags, head POS, and SRL

tags. This comprehensive dataset served as the foundation for our experiments, providing contextual information necessary for training neural networks and conducting experiments on Semantic Role Labeling (SRL).

Example- Dataset1 :

Word	Chunk	Postposition	Dependency-Head	Dependency	Is_Arg	SRL	Predicate	Head-POS
उल्लेखनीय	JJP	0	VGF	k1s	1	ARG2-ATR	VGF.1	adj
है	VGF	है	0	root	0			v
कि	CCP	0	NULL__NP	rs	0			avy
अक्तूबर	NP	0_को	VGNF	k7t	1	ARGM-TMP	VGNF	n
आए	VGNF	या	NP2	nmod__k1inv	0			v
भूकंप	NP2	0_के_बाद	VGF2	k7t	1	ARGM-TMP	VGF2	n
जिंदा	JJP2	0	VGNF2	pof	0			adj
बचे	VGNF2	या	NP3	nmod__k1inv	0			v
लोगों	NP3	0_में	VGF2	k7p	1	ARGM-LOC	VGF2	n

Example Dataset2 :

Sentence	Dependency	SRLs
सदियों मानव मन सवाल रहा कि क्या अंतरिक्ष हम अकेले हैं या ब्रह्मांड ग्रह जीवन स्वरूप मौजूद है2	k7t r6 k7 k1 root rs vmod k7p k1 k1s ccof ccof r6 k7p k1 k7 k1s ccof	ARGM-TMP NO_SRL ARG2-ATR ARG1 NO_SRL NO_SRL ARGM-DIS ARGM-LOC ARG1 ARG2-ATR NO_SRL NO_SRL NO_SRL ARGM-LOC ARG1 ARGM-MNR ARG2-ATR NO_SRL
अभी वैज्ञानिकों सवाल जवाब मिला लेकिन ग्रहों खोज वैज्ञानिकों2 उम्मीदें बढ़ा	k7t k4 r6 k2 ccof root r6 k1 r6 k2 ccof	ARGM-TMP ARG0 NO_SRL ARG1 NO_SRL NO_SRL NO_SRL ARG0 NO_SRL ARG1 NO_SRL
पृथ्वी ग्रहों जीवन उम्मीद है	nmod k7p r6 k1 root	NO_SRL ARGM-LOC NO_SRL ARG1 NO_SRL
ग्रह पृथ्वी ज्यादा बड़े हैं	k1 k1u jjmod k1s root	ARG1 ARGM-MNR NO_SRL ARG2-ATR NO_SRL

2. **Statistical models :** Our approach involved employing the Linear Support Vector Classifier (LinearSVC) from scikit-learn to train statistical models for Semantic Role Labeling (SRL). This classifier is known for its effectiveness in handling high-dimensional data and is commonly used for text classification tasks.

   We experimented with three different sets of input features, each capturing different aspects of the input data:

   1. **Model1**: This set includes features such as the word, whether it's an argument, the predicate, and the postposition.
   2. **Model2**: These features include the word, postposition, whether it's an argument, the predicate, and the head-POS.
   3. **Model3**: This set comprises features such as the word, dependency, postposition, whether it's an argument, the predicate, and the head-POS.

   After encoding the features and target labels using LabelEncoder, we adopted a standard approach of splitting the data into training and testing sets. The training data was used to fit the model, while the testing data was used to evaluate its performance.

   We then trained each model using the LinearSVC classifier and assessed its performance using standard evaluation metrics such as accuracy, precision, recall, and F1-score. These metrics provided insights into the effectiveness of each model in classifying semantic roles.

3. **Neural network models :** In our exploration of neural network models, we experimented with various architectures and embeddings to enhance Semantic Role Labeling (SRL) performance.

   1. FastText Embeddings + BiLSTM Classifier: We began by incorporating FastText embeddings into a Bidirectional Long Short-Term Memory (BiLSTM) classifier. FastText embeddings offer a computationally efficient means of representing words in vector space. The BiLSTM layer captures contextual information from both past and future tokens, enabling the model to understand the sequential nature of language.

   2. FastText Embeddings + Dependency Relation + BiLSTM Classifier: Building upon the previous model, we augmented the input features with

dependency relations extracted from the dataset. Dependency relations provide valuable syntactic and semantic information about the relationships between words in a sentence. By incorporating these relations alongside FastText embeddings, we aimed to improve the model's ability to discern the roles of different participants in events.

3. Indic-Bert + MLP Classifier: Next, we leveraged the powerful representations learned by the Indic-Bert model, a transformer-based architecture specifically trained for Indian languages. We utilized the AutoModel and AutoTokenizer classes from the Hugging Face Transformers library to instantiate the Indic-Bert model and tokenizer. The embeddings obtained from Indic-Bert were fed into a Multilayer Perceptron (MLP) classifier for SRL classification. This architecture allowed us to benefit from the contextual understanding encoded by Indic-Bert while maintaining flexibility in the classification layer.

4. Indic-Bert + BiLSTM Classifier: Lastly, we combined Indic-Bert embeddings with a BiLSTM classifier. This architecture aimed to capitalize on the strengths of both transformer-based models and recurrent neural networks. The Indic-Bert embeddings provided rich contextual information, while the BiLSTM layer offered additional sequential modeling capabilities. By fusing these components, we sought to create a robust SRL system capable of capturing nuanced relationships between words in Indian languages.

5. Indic-Bert + Dependency Relation + BiLSTM Classifier:
Expanding on our exploration, we introduced the integration of Indic-Bert embeddings with dependency relations into a Bidirectional Long Short-Term Memory (BiLSTM) classifier. Indic-Bert embeddings capture rich contextual information specific to Indian languages, offering a robust foundation for semantic understanding. By incorporating dependency relations extracted from the dataset alongside Indic-Bert embeddings, we aimed to enrich the model's understanding of syntactic and semantic relationships within sentences. The BiLSTM layer further enhanced the model's ability to capture sequential dependencies, complementing the contextual knowledge encoded by Indic-Bert. Through this fusion of components, we aimed to develop a comprehensive SRL system capable of accurately identifying semantic role labels in Indian languages.

Each model was trained using the Adam optimizer and the Cross-Entropy Loss function. We conducted multiple epochs of training to optimize model parameters and minimize loss. Subsequently, the trained models were evaluated on a

separate test dataset to assess their performance in predicting semantic role labels accurately.

5. Evaluation :

Baseline Model: In our baseline model, we utilized a straightforward approach by considering only the most fundamental features for Semantic Role Labeling (SRL). These features include:

- **Word** : The actual word token extracted from the sentence.
- **POS Tag**: The part-of-speech tag assigned to each word, indicating its grammatical category.
- **Is Argument:** A binary indicator representing whether the word is **an** argument of the predicate or not.

These features serve as the foundation for our classification task, providing essential information about each word's identity, grammatical function, and role within the sentence.

Statistical models :

Experimental Results :

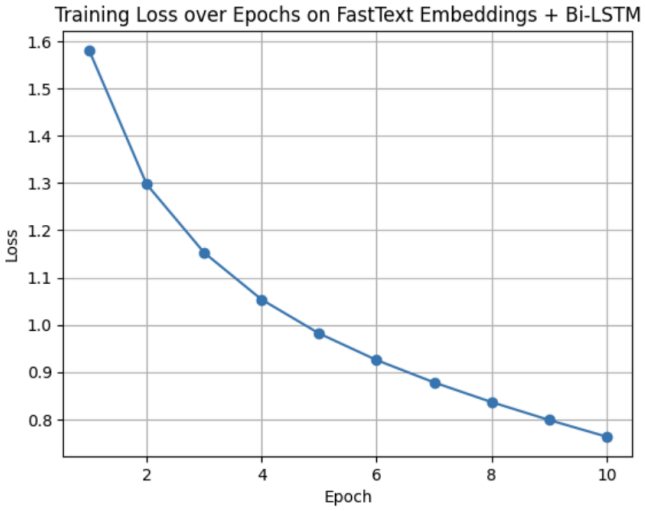
Model	Features	Accuracy	Precision	Recall	F1 Score
Model 1	Baseline	76.94%	69.27	76.94	71.95
Model 2	Baseline + 'Head POS Tag'	77.22%	73.41	77.22	72.75
Model 3	Baseline + 'Head POS Tag' + 'Dependency'	81.27%	77.53	81.27	78.00

Inference Results :

Predicted Word	Semantic Role Label
ग्रह	nan
पार्टी	ARG1
कदम	nan
नियमों	ARG1
भारत	nan
गोले	ARG1
उन्होंने	ARG1
डांस	ARG1
भूमिका	nan
करने	ARGM-MNR

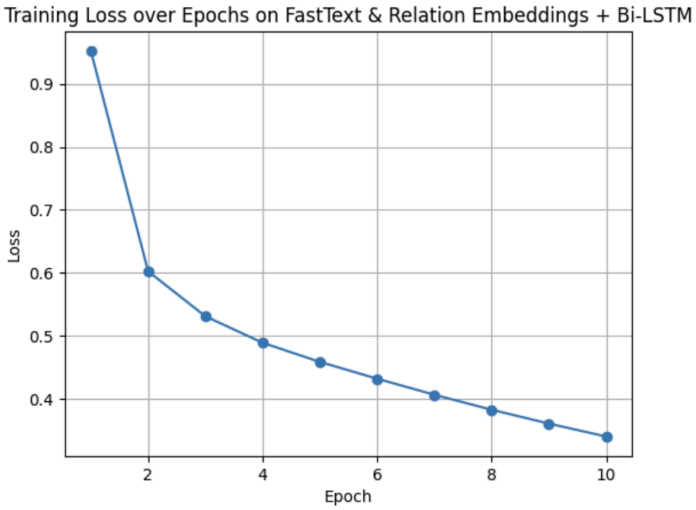
Neural network models:

1. FastText Emb + BiLSTM Classifier :



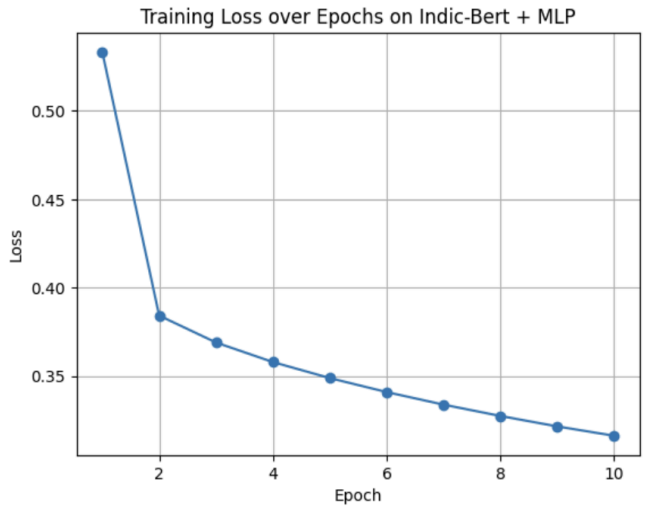
Test Accuracy : 70.41 %

2. FastText Emb + Dependency Relation + BiLSTM Classifier :



Test Accuracy : 86.86%

3. Indic-Bert + MLP Classifier :

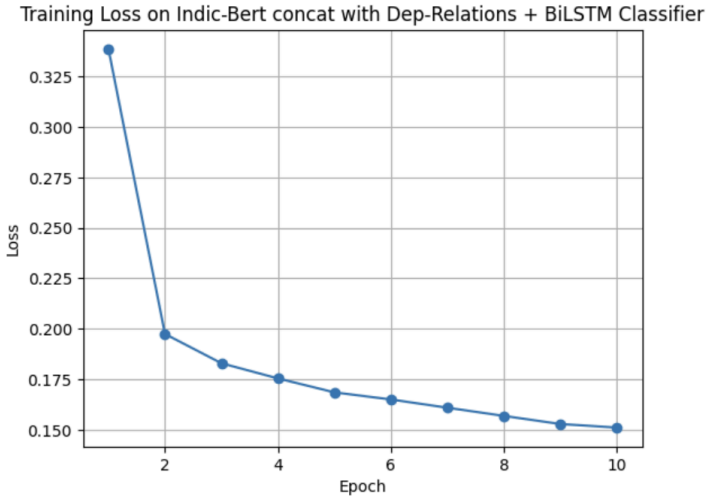


Test Accuracy : 89.81%

4. Indic-Bert + Bi-Lstm Classifier :

Test Accuracy : 90.02%

5. Indic-Bert + Dependency Relation + Bi-Lstm Classifier:



Test Accuracy : 95.34%

Final Results :

Model	Accuracy	Precision	Recall	F1 Score
Fast-Text Embedding with Bi-LSTM	0.7041	0.6400	0.7000	0.6600
Fast-Text Embedding and Dependency Relation with Bi-LSTM	0.8686	0.8600	0.8700	0.8600
BERT with MLP	0.8987	0.8524	0.8987	0.8681
BERT with Bi-LSTM	0.9007	0.8670	0.9007	0.8823
BERT and Dependency Relation with Bi-LSTM	0.9534	0.9510	0.9534	0.9516

6. Future Work :

Exploring Different BERT Layers for Embeddings:

Diving into the various layers of BERT offers an opportunity to improve how semantic information is represented in semantic role labeling (SRL) tasks. BERT is built with multiple layers that process input sentences to generate contextual embeddings.

Testing different layers helps identify which ones capture the most important semantic features for SRL. Typically, layers closer to the input focus on fine syntactic details, while deeper layers capture broader, higher-level meanings.

By carefully choosing BERT layers to extract embeddings, SRL models can balance computational efficiency with semantic richness, leading to more streamlined and effective model designs.

7. Code : Codes and saved models.

8. References :

1. Gupta, A., & Shrivastava, M. (2018). Enhancing Semantic Role Labeling in Hindi and Urdu. In *Proceedings of the LREC 2018 Workshop on Representation Learning for NLP* (pp. 28-32). Retrieved from Link

2. Anwar, M., & Sharma, D. (2016). Towards Building Semantic Role Labeler for Indian Languages. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16) (pp. 4588–4595). Portorož, Slovenia: European Language Resources Association (ELRA). Retrieved from Link

3. Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2016). Enriching Word Vectors with Subword Information. Facebook AI Research. Retrieved from Link