Semantic Role Labeling

Team Name: **Semantic Sentinels**
Team Number: **55**
1. D Priyanka - 2023814003
2. S Monica - 2023802005
3. CV Thirumala Kumar - 2023702020

Introduction

Semantic Role Labeling (SRL) is the task of assigning roles to words/phrases in relation to the main verb, such as "agent", "patient", "instrument", etc.
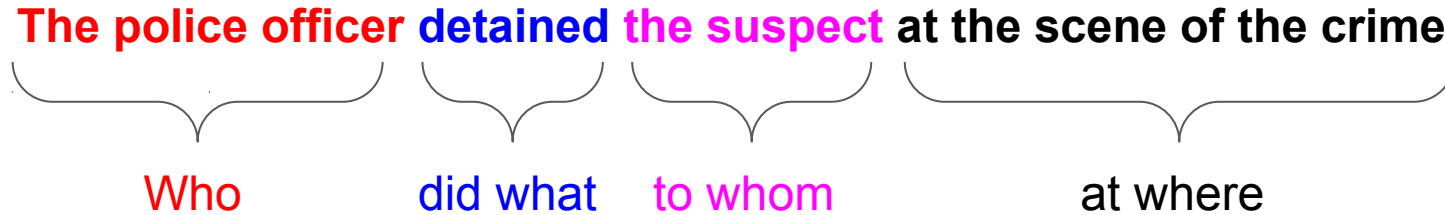
The police officer detained the suspect at the scene of the crime

Who   did what   to whom   at where

Fig. Pictorial depiction of SRL

Core Semantic Role Labels

Table 1: Core Semantic role labels

| Tags	Core Arguments	Description
ArgA	AGENT	The volitional causer of an event
ArgT	THEME	The participant most directly affected by an event
ArgEx	EXPERIENCER	The experiencer of an event
ArgB	BENEFICIARY	The beneficiary of an event
ArgR	RECIPIANT	Receive something (whether good or bad) in an event
ArgI	INSTRUMENT	An instrument used in an event
ArgLOC	LOCATION	A locative or path prepositional phrase introduces an underspecified location.
ArgTOP	TOPIC	Conversation or message transfer is dealt with using communication verbs
ArgG	GOAL	Verb's destination or goal argument that indicates the motion's endpoint.
ArgS	SOURCE	arguments that can be thought of as a source or beginning point for the verb's event

Example
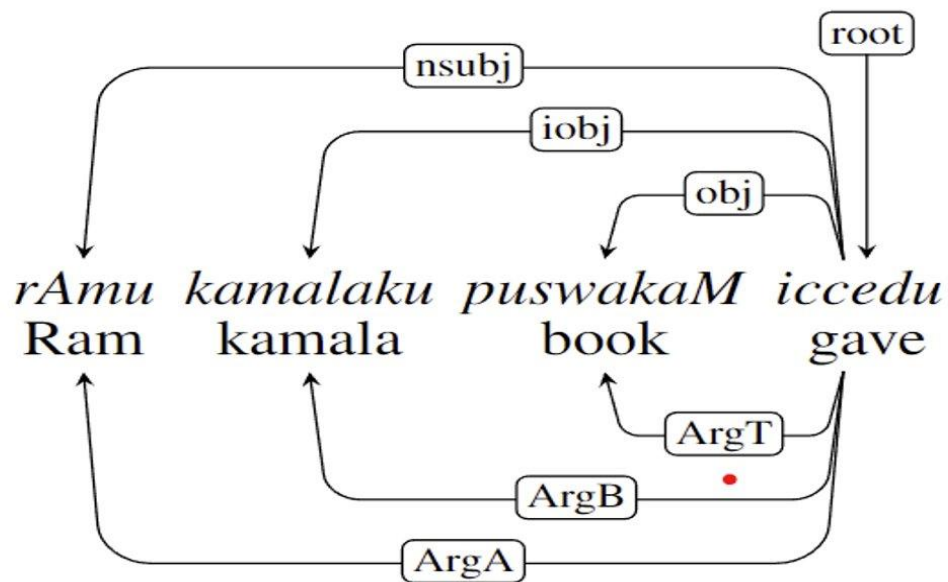
Figure 1: 'Ram gave a book to Kamala'

Existing approaches

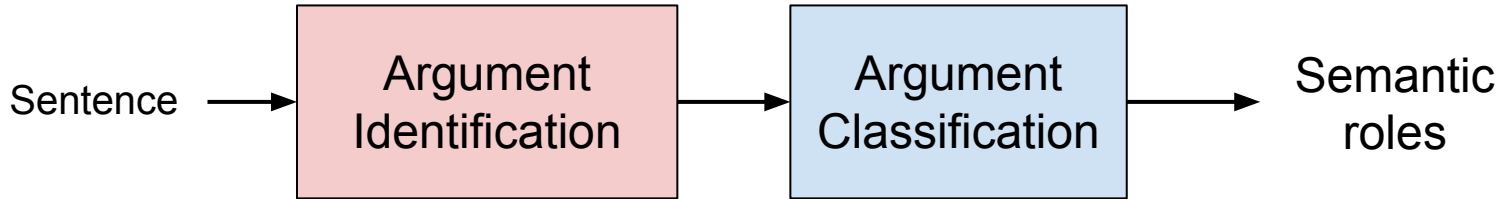Sentence → **Argument Identification** → **Argument Classification** → Semantic roles
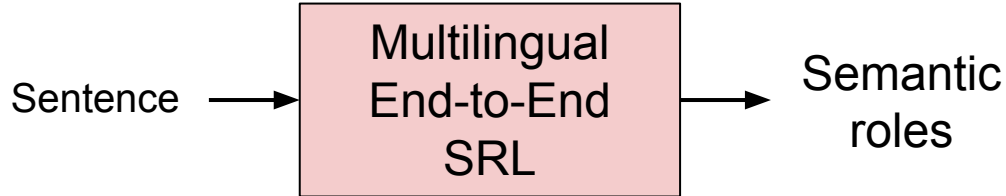
Fig. Cascaded approach

Drawbacks:
1. Requires additional information such as PoS tags, dependency relation etc.
2. Not feasible for practical usage
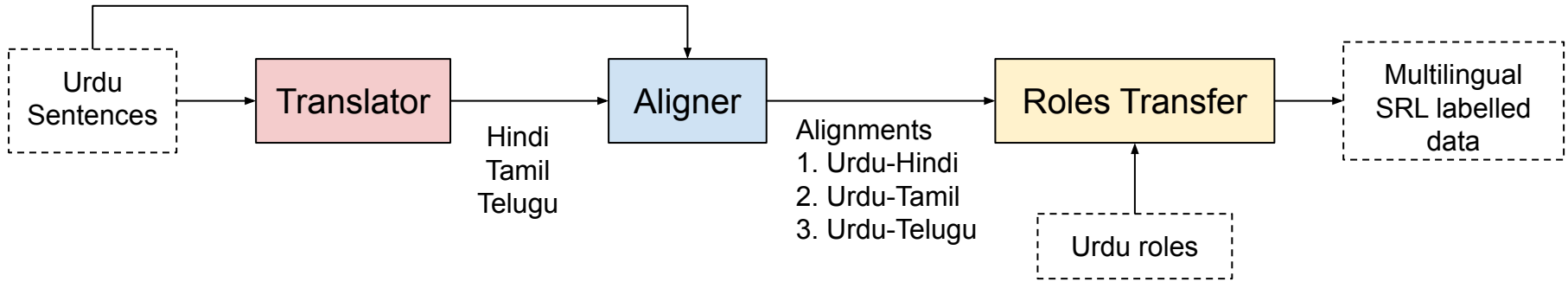3. Not feasible for multiple languages

Proposed approach

Sentence → Multilingual End-to-End SRL → Semantic roles

Challenges:
1. Lack of data for Deep learning approach
2. Lack of Multi-lingual SRL data

Available Data:
1. Urdu SRL dataset

Data Augmentation Pipeline

Non-SRL Tags:
1. NAH (Not-A-Head)
2. NAA (Not-An-Argument)
3. Predicate

Experimental setup

Fine-tuning 2 bert based models for SRL as a token-classification task.
1. Indic-Bert (AI4Bharath) - 24 Languages
2. Multilingual-Bert (Google) - 104 Languages

Model Name	Base Model	Description
Indic-bert	Indic-bert	labelling only first subtoken of the word
Indic-bert -lat		labelling all sub tokens of the word
multilingual-bert	multilingual bert	labelling only first subtoken of the word
multilingual-bert-lat		labelling all sub tokens of the word

Training Configuration:
1. Batch size - 16 per 2 device (2 GPUs used)
2. Learning rate 0.00005
3. Training epochs 20 with Early Stopping with patience 5

Results

Baseline SVM vs Proposed approach

Model	Weighted F1		Macro F1	
	Argument Identification	Argument Classification	Argument Identification	Argument Classification
Indic-bert	0.60	0.54	0.59	0.21
Indic-bert -lat	0.64	0.60	0.628	0.22
multilingual-bert	**0.71**	**0.68**	**0.70**	**0.34**
multilingual-bert-lat	0.70	0.67	0.69	0.31
SVM	**0.77**	**0.64**	**0.77**	**0.32**

Observation: Proposed approach showed similar performance as baseline

Results

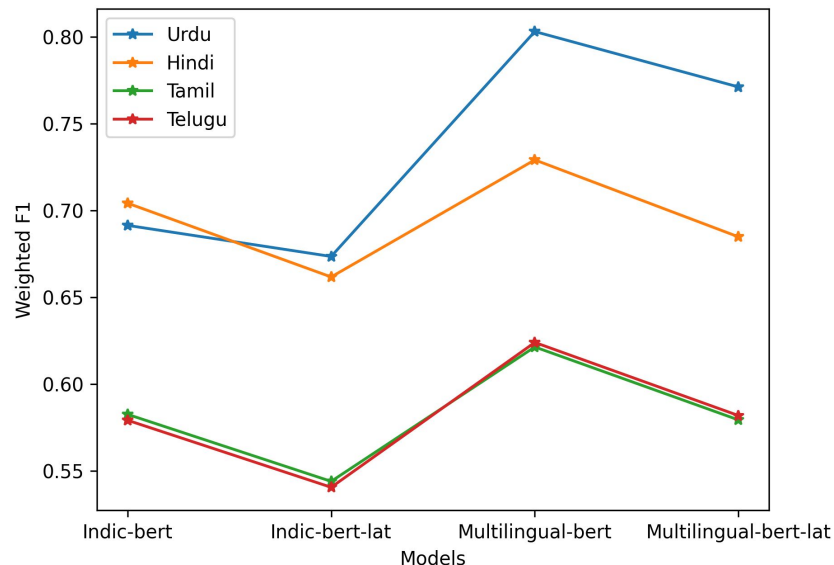## Language wise overall performance comparison
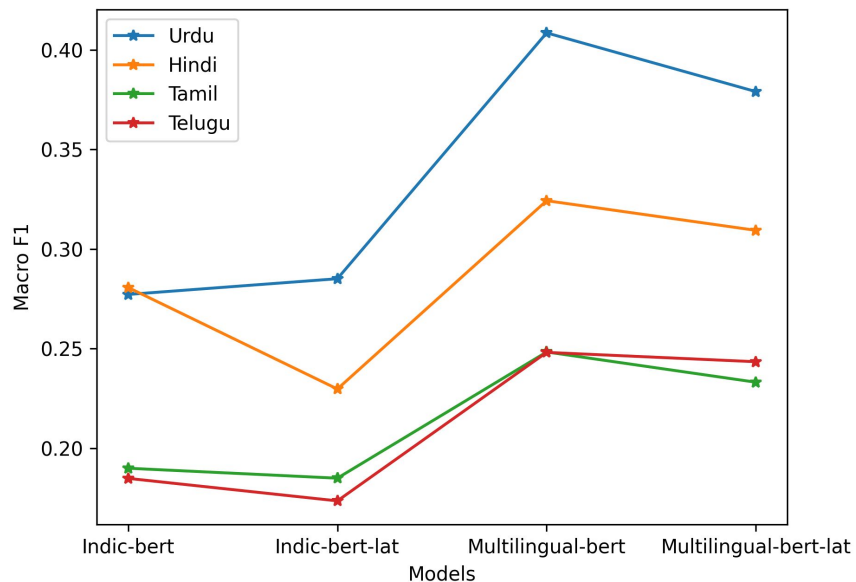
Observations:
1. Urdu and Hindi showed similar performance
2. Tamil and Telugu Showed similar performance

Possible reasons:
1. Urdu-Hindi and Tamil-Telugu belongs to different language families

Conclusion

1. Multilingual End-to-End SRL method is proposed using multilingual data augmentation from a single language labelled data.
2. Proposed approach performing similar to the baseline without needing additional information such as PoS tags, dependency relations etc.
3. Language characteristic based trend is observed in the performance.
4. Extending this approach for more languages, observing source language impact on target is the future scope of this project

Thank You