# Template Extraction from unstructured Wikipedia text using NLP techniques.

1 st Sameer Ahmed Btech-IT) IIIT Allahabad) UP, India
iit2020053@iiita.ac.in

# Index

1)Abstract

2) Problem Statement

3)Proposed Solution

4)Dataset Description

5)Future Scope

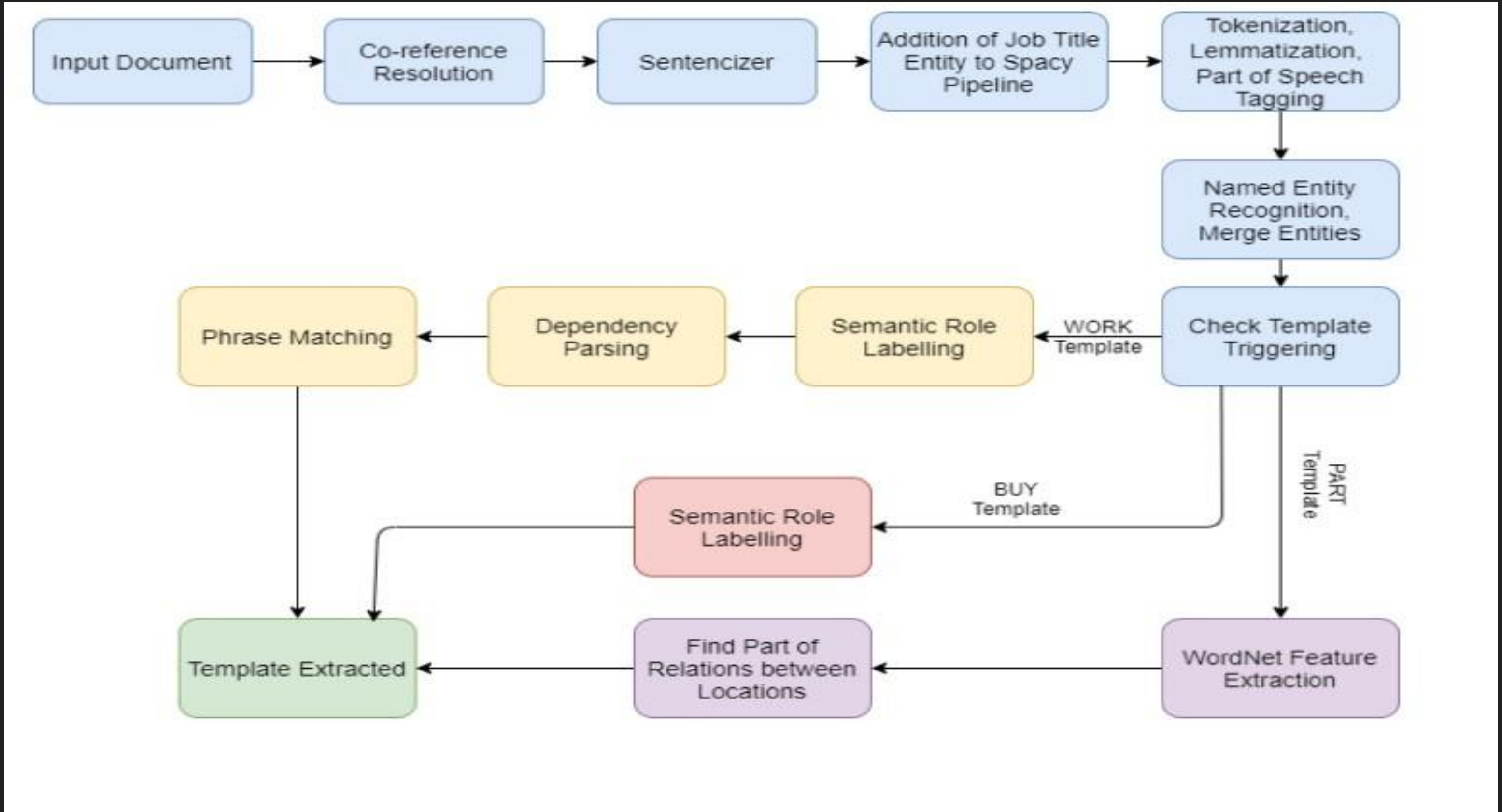
# Abstract

This document describes a project on information extraction using template filling in natural language processing (NLP) and linguistics. The project involves extracting information from unstructured textual data such as Wikipedia articles and filling three given templates for buying events, job title relations, and part-of relations. The proposed solution for the project focuses on a rule-based approach, which involves defining a set of syntactical and grammatical rules for natural language and using various techniques such as named-entity recognition, coreference resolution, semantic-role labeling, dependency parsing, phrase matching, and wordnet features to extract information and fill the templates. The document provides an example sentence for each template and outlines the steps involved in extracting the relevant information to fill the templates.

# Problem Description

The task is to perform information extraction using template (slot) filling on unstructured textual data, such as Wikipedia articles, by applying various NLP techniques. The goal is to extract information and fill three given templates related to buying events, job title relations, and part-of relations between locations. 30 Wikipedia articles are provided, 10 each related to organizations, persons, and locations.

# Proposed Solution


Input Document
Co-reference Resolution
Sentencizer
Addition of Job Title Entity to Spacy Pipeline
Tokenization, Lemmatization, Part of Speech Tagging
Named Entity Recognition, Merge Entities
Check Template Triggering
WORK Template
Semantic Role Labelling
Dependency Parsing
Phrase Matching
BUY Template
Semantic Role Labelling
PART Template
WordNet Feature Extraction
Find Part of Relations between Locations
Template Extracted

# Dataset Description-

We were provided with 30 Wikipedia Articles:

- 10 articles related to Organizations
- 10 articles related to Persons
- 10 articles related to Locations

And we were expected to extract information from the above mentioned articles and fill the given 3 templates:

# Future Scope

- Instead of using default model for NER, SRL and Coreference provided by Spacy and AllenNlp, on top of it, we could train a model to increase our performance.
- We could further find instances where our proposed solution fails and provide a workaround for that

Thank you