Crowdsourcing Approach in Statistical Machine Translation

Project Title

Development of Multilingual Resources and Technologies for Indian Languages

(a collaborative project between IIT Bombay and Xerox)

Presented By:- Rajen Chatterjee (Research Engineer, IIT Bombay)

Team Members:-Abhijit Mishra (Research Scholar, IIT Bombay)

Anoop Kunchukuttan (Research Scholar, IIT Bombay)

Outline

- What is Crowdsourcing?
- Role of Crowdsourcing in SMT
- How to manage crowd?
- Basic terminologies
- Need for automation?
- Crowdsourcing Platform Architecture
- Crowdsourcing Engine
- Workflow
- Different Stages
- Pipelining and Quality Control
- A working example

What is Crowdsourcing?

- **Crowdsourcing** is a process that involves outsourcing tasks to a distributed group of people.
  - Example: Recaptcha used in Gmail account creation.
- This process can occur both online and offline.
- The difference between crowdsourcing and ordinary outsourcing is that a task or problem is outsourced to an undefined public rather than a specific body.

Role of crowdsourcing in SMT

- Crowdsourcing approach is being used for development of Parallel Corpora.
- Not just development of parallel Corpora but development of qualitative parallel Corpora.
- Ensuring quality is the most challenging task in Crowdsourcing.
- Crowd Tasks:
  - Phrase Translation
  - Phrase Validation
  - Phrase Re-Combination

How to manage crowd?

- Third party service provider like Amazon Mechanical Turk, CrowdFlower.
- Amazon Mechanical Turk is a web service that provides an on-demand, scalable, human workforce to complete jobs that humans can do better than computers.
- Sample Business Cases:
  - Translation
  - Image Tagging
  - Data Verification
  - Information Gathering
- Service Highlights:
  - On-Demand Workforce
  - Quality Management Tools
  - Lower Cost Structure

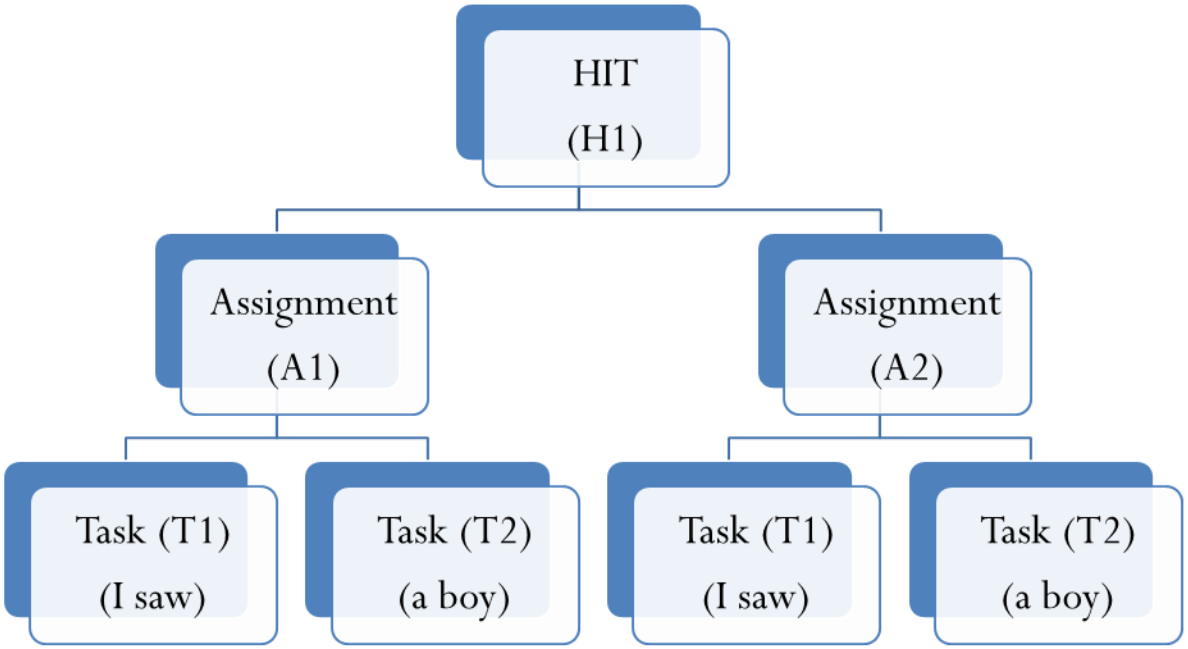
Example..

- Consider the following small sentence

  I saw a boy.

- Our objective is to get the Hindi translation of the whole sentence.
- In first phase, will divide it into two phrases viz. “I saw” and “a boy” and will obtain translations of the corresponding phrases from the crowd.
- Second phase will ensure whether the quality of the target phrase is good or not.
- If the quality is good, we’ll recombine the phrases in the final phase to get the whole target sentence.

Basic Terminologies



- Task:
  - Task is the atomic unit of work.
- Assignment:
  - Several task are combined into one Assignment.
- HIT (Human Intelligence Task):
  - Several Assignment are combined into one HIT.

Need for Automation?

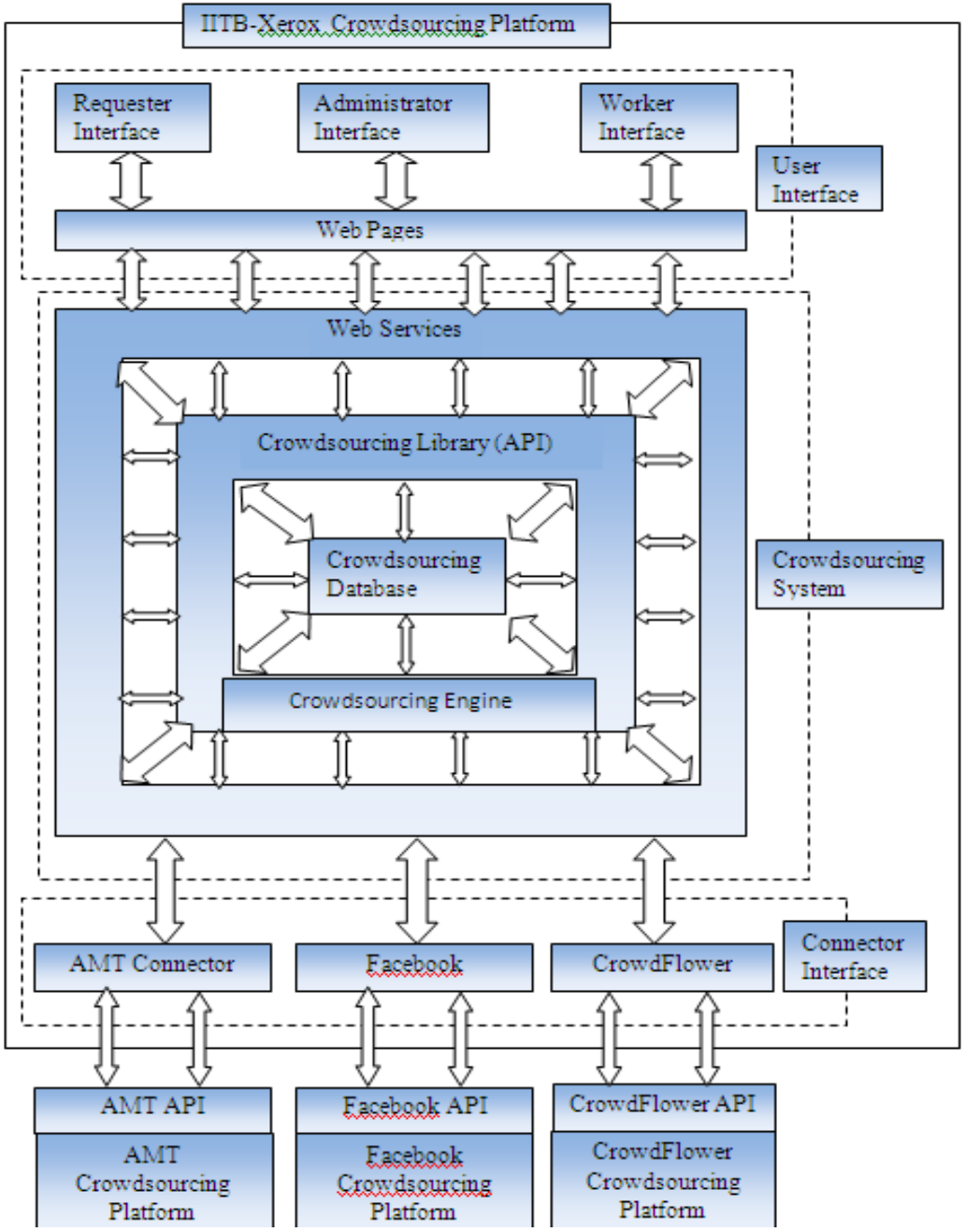
- Manually managing Crowdsourcing task is very tedious specially in case of Translation.
- So there is a need to automate the whole crowd translation process.
- Automation will help to get the following things done:
  - Get the Translation of each source phrase
  - Validate the translated phrase.
  - Re-combine the translated phrase into sentence.
- This brings up the motivation for developing Crowdsourcing Platform.

Crowdsourcing Platform

- Goal of this platform is to develop a generic task independent platform but focusing more on Translation task.

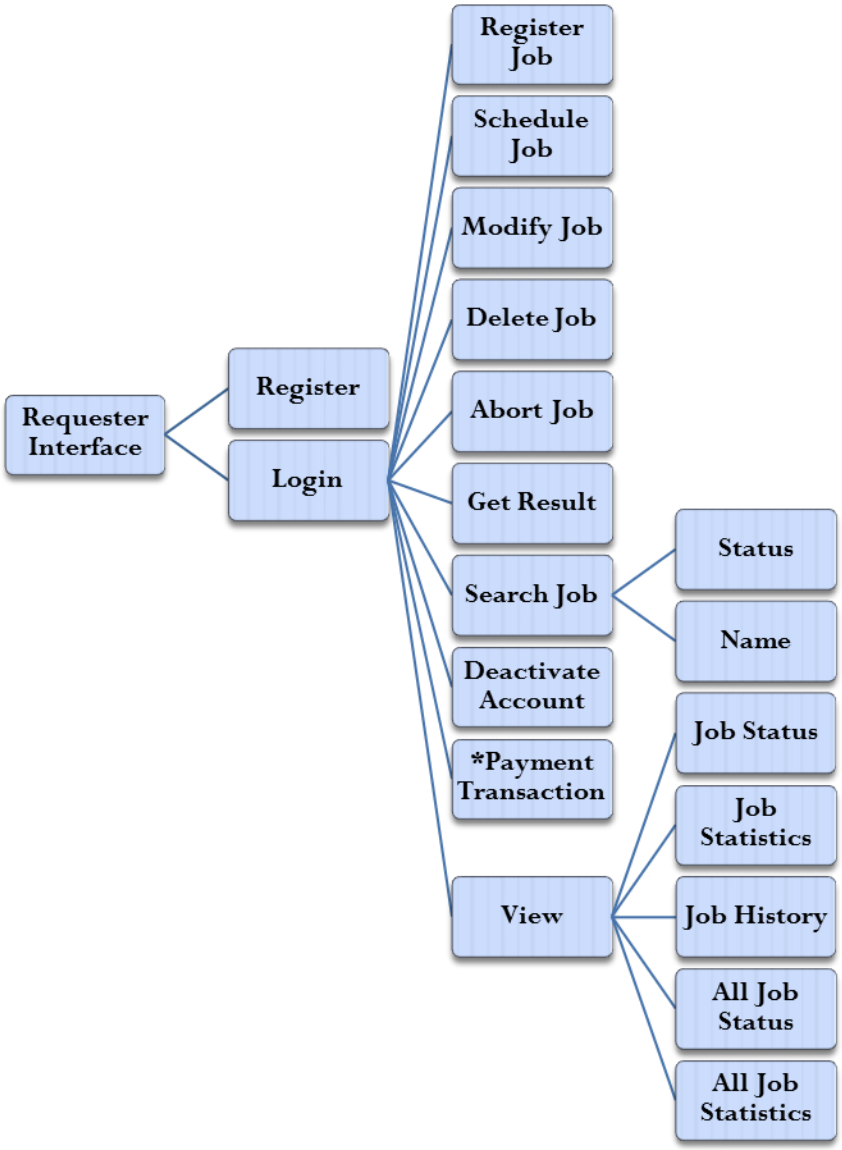
- From translation point of view our goal is to develop language independent and domain independent parallel corpora.

ARCHITECTURE



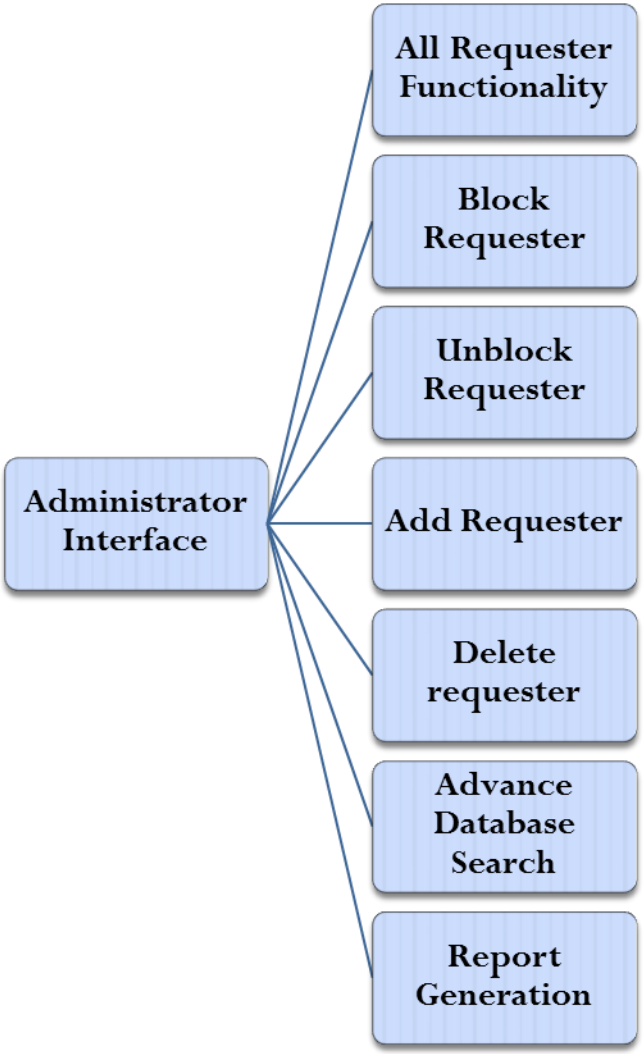
Requester

- Requester is the one who requests for getting a translation of a given document.



Administrator

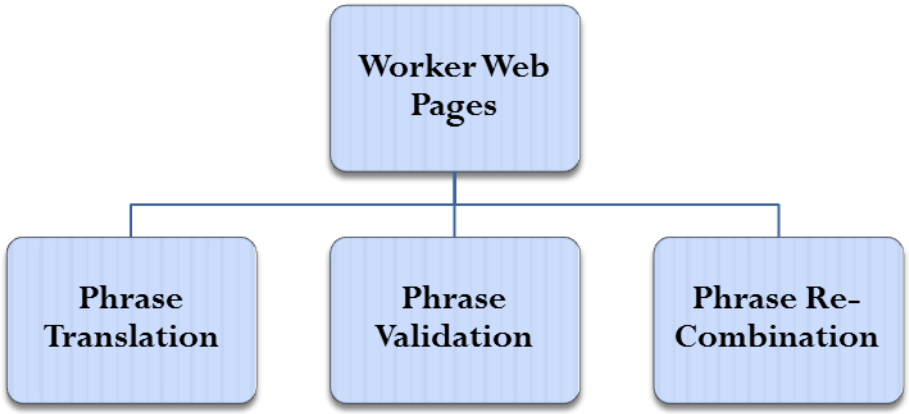
- Administrator will have all rights to perform all tasks on behalf of a requester



Worker

- Workers are the member of crowd itself who are actually involved in solving the task. Workers will be connected to Crowdsourcing worker web page indirectly via other Crowdsourcing platform like AMT.

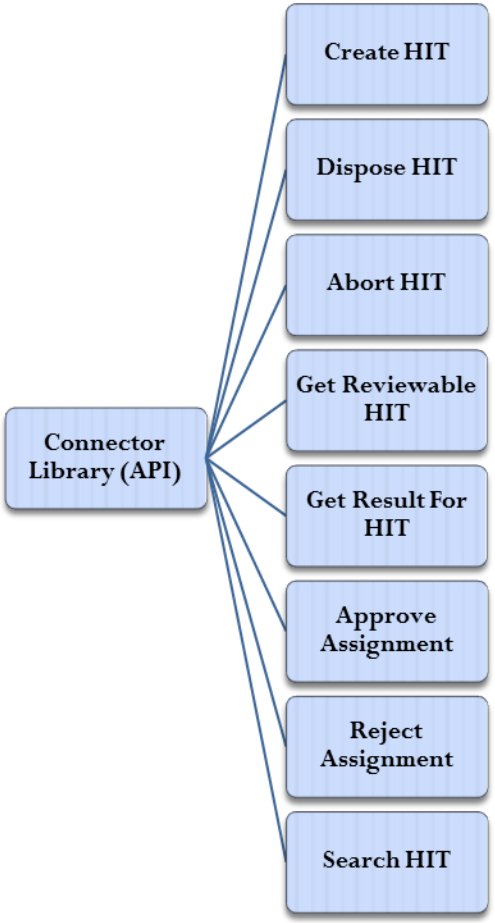
- Workers will be provided with any one of the following web pages for translation job.
  - Phrase Translation
  - Phrase Validation
  - Phrase Re-Combination



Platform Connector

- Platform connector behaves as an interface between IITB-Xerox Crowdsourcing system and other Crowdsourcing platform.

- Each connector will have its own API functionality which will interact with connector Crowdsourcing platform.



Platform Connector

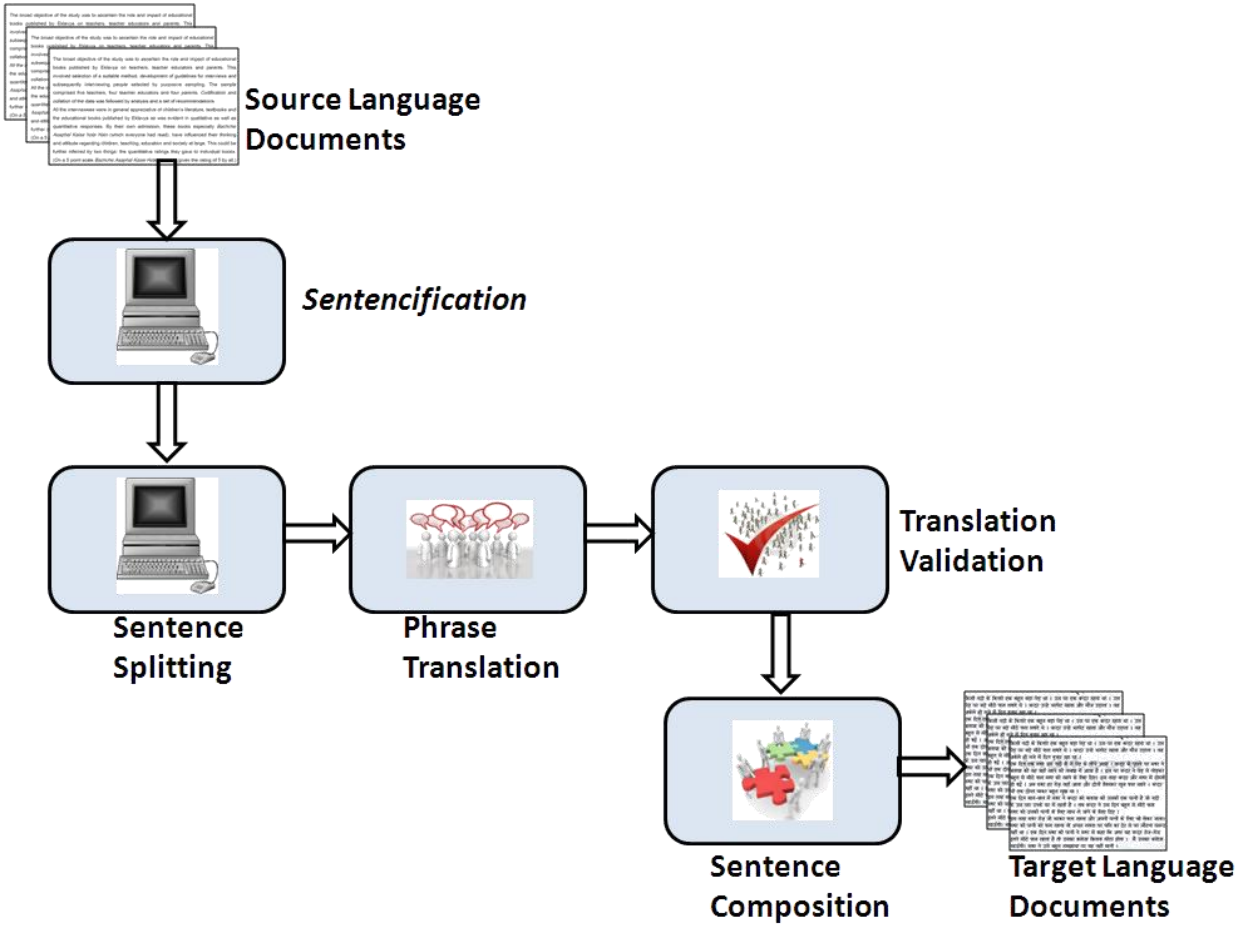
- Following are the role of a connector:
  - Pulling And Pushing Mechanism:
    - Pulling data from IITB-Xerox Crowdsourcing system and pushing into AMT.
    - Pulling result from AMT and pushing into IITB-Xerox Crowdsourcing system.
  - Monitoring:
    - Periodically monitoring IITB-Xerox Crowdsourcing system to check any status change of a HIT
    - Periodically monitoring AMT Crowdsourcing platform to check for the result of HIT.

Crowdsourcing Engine

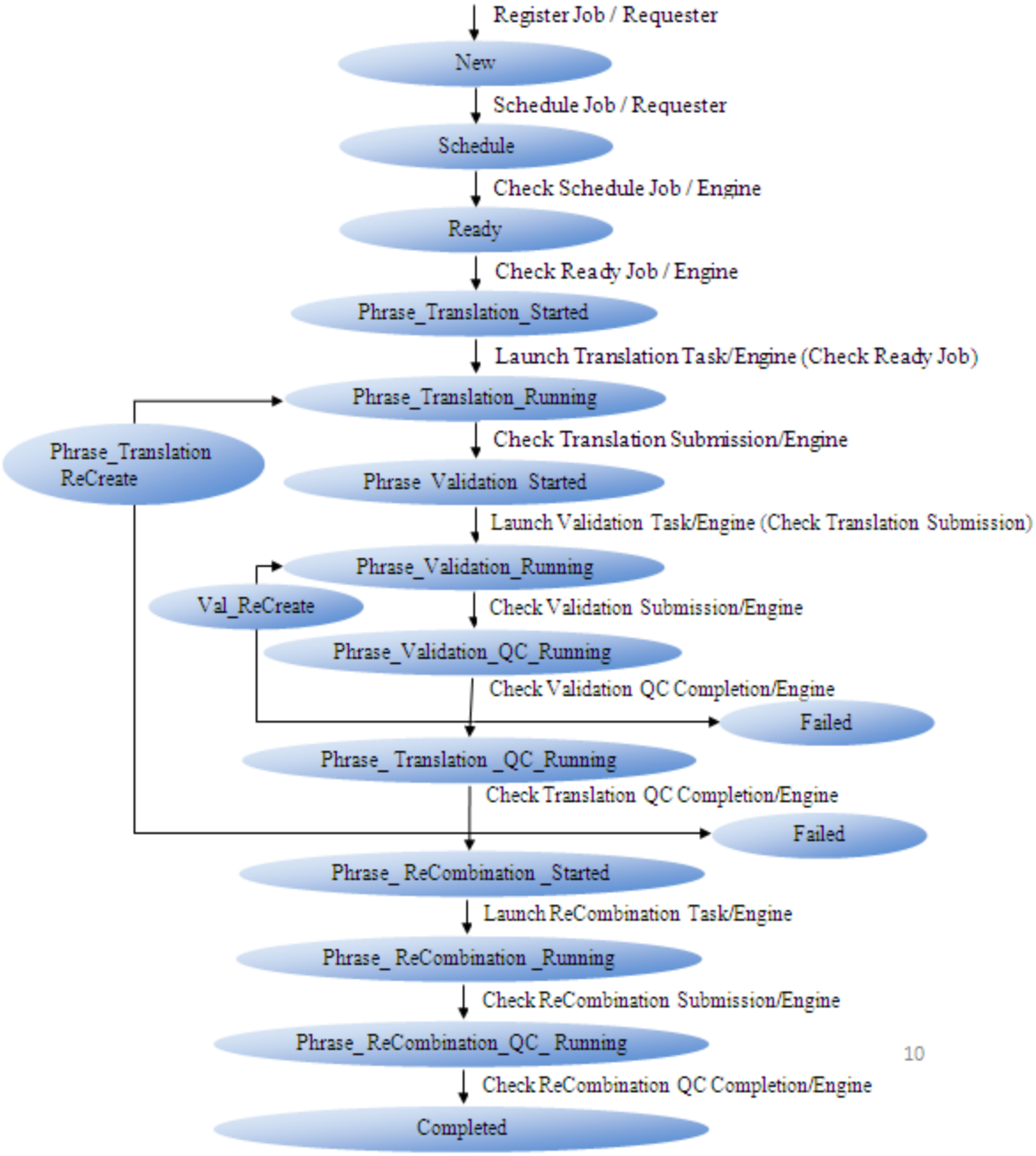
- Crowdsourcing Engine is the heart and soul of the Crowdsourcing Platform.

- Following are the responsibility of the engine:
  - Monitoring entire life cycle of a job.
  - Managing the Pipelining architecture.
  - Ensuring Quality Control Mechanism.
  - Update the HIT status whenever Job status changes.

Workflow (A Naïve view)

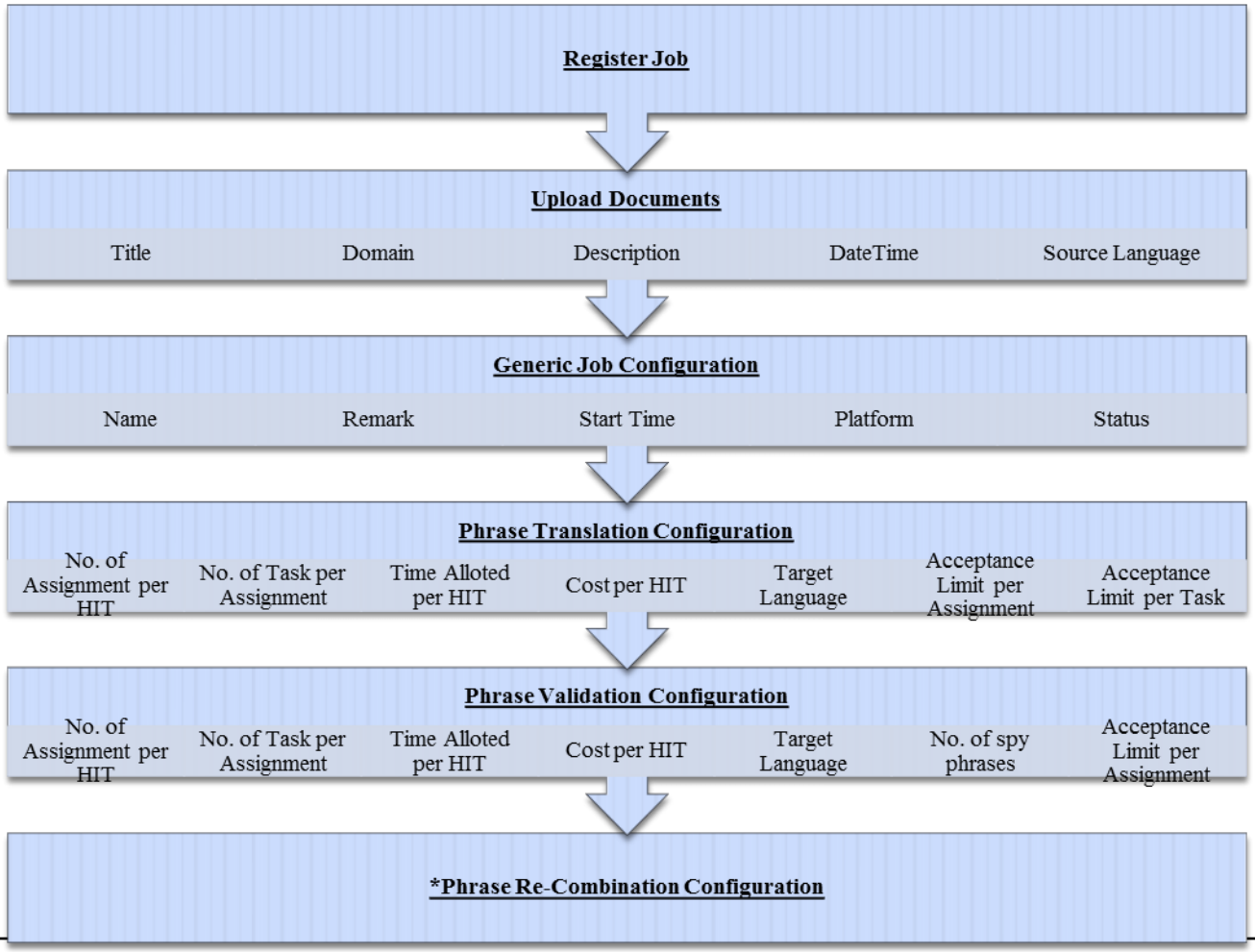


WORKFLOW

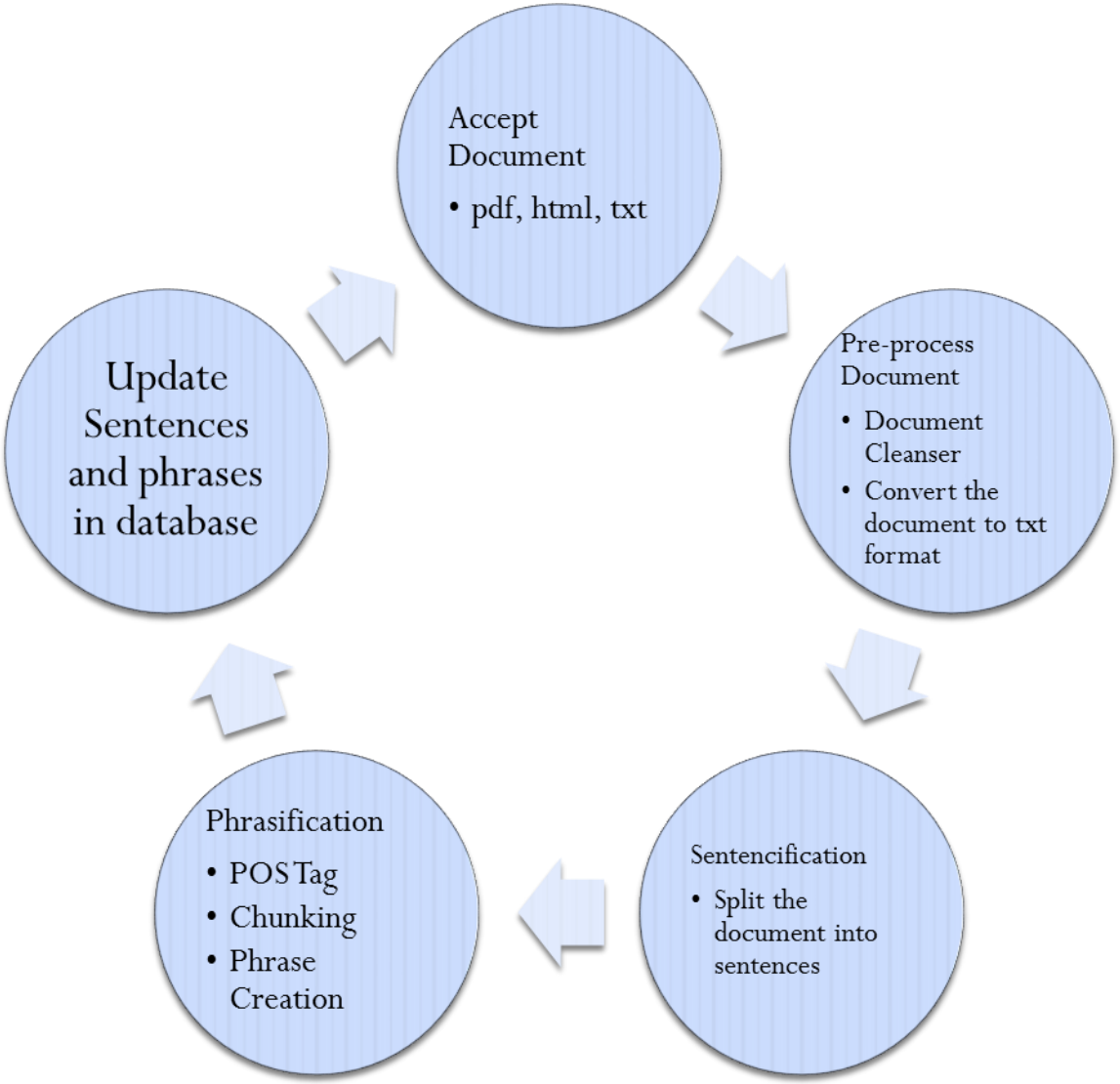

Register Job / Requester
New
Schedule Job / Requester
Schedule
Check Schedule Job / Engine
Ready
Check Ready Job / Engine
Phrase_Translation_Started
Launch Translation Task/Engine (Check Ready Job)
Phrase_Translation_Running
Phrase_Translation ReCreate
Check Translation Submission/Engine
Phrase Validation Started
Launch Validation Task/Engine (Check Translation Submission)
Phrase_Validation_Running
Val_ReCreate
Check Validation Submission/Engine
Phrase_Validation_QC_Running
Check Validation QC Completion/Engine
Failed
Phrase_ Translation _QC_Running
Check Translation QC Completion/Engine
Failed
Phrase_ ReCombination _Started
Launch ReCombination Task/Engine
Phrase_ ReCombination _Running
Check ReCombination Submission/Engine
Phrase_ReCombination_QC_ Running
Check ReCombination QC Completion/Engine
Completed

Register job

- Register job will only create a job for execution i.e. it will register all job configuration parameters in the database and upload the documents for translation



Document processor



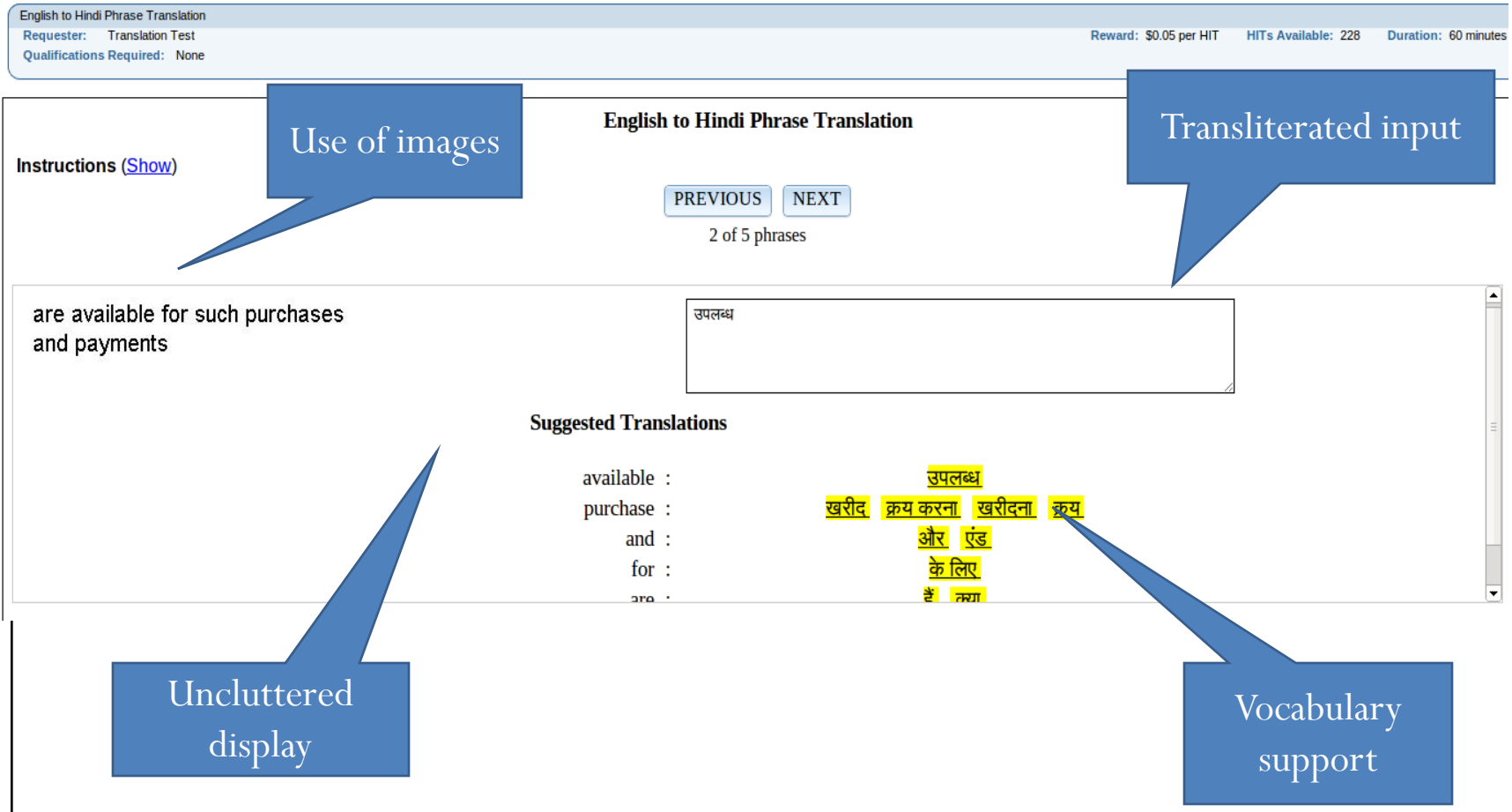
Pipelining & Quality Control

- Pipelining ensure that different stages in entire translation process runs smoothly.

- Quality Control ensures that we get the translated phrase of reasonable quality.

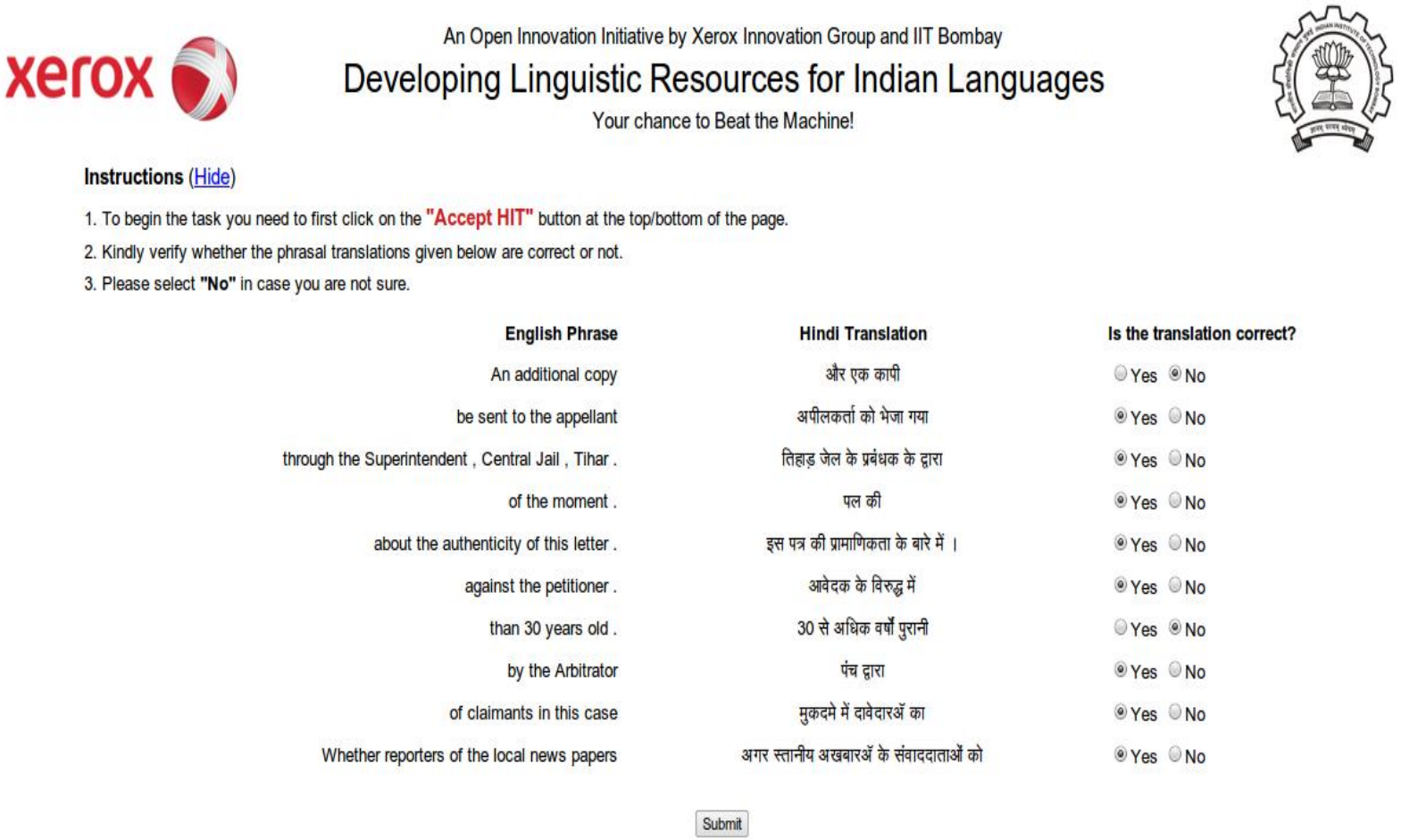
Phrase Translation

- In Phrase Translation, given a source phrase the worker is required to give the translated phrase.



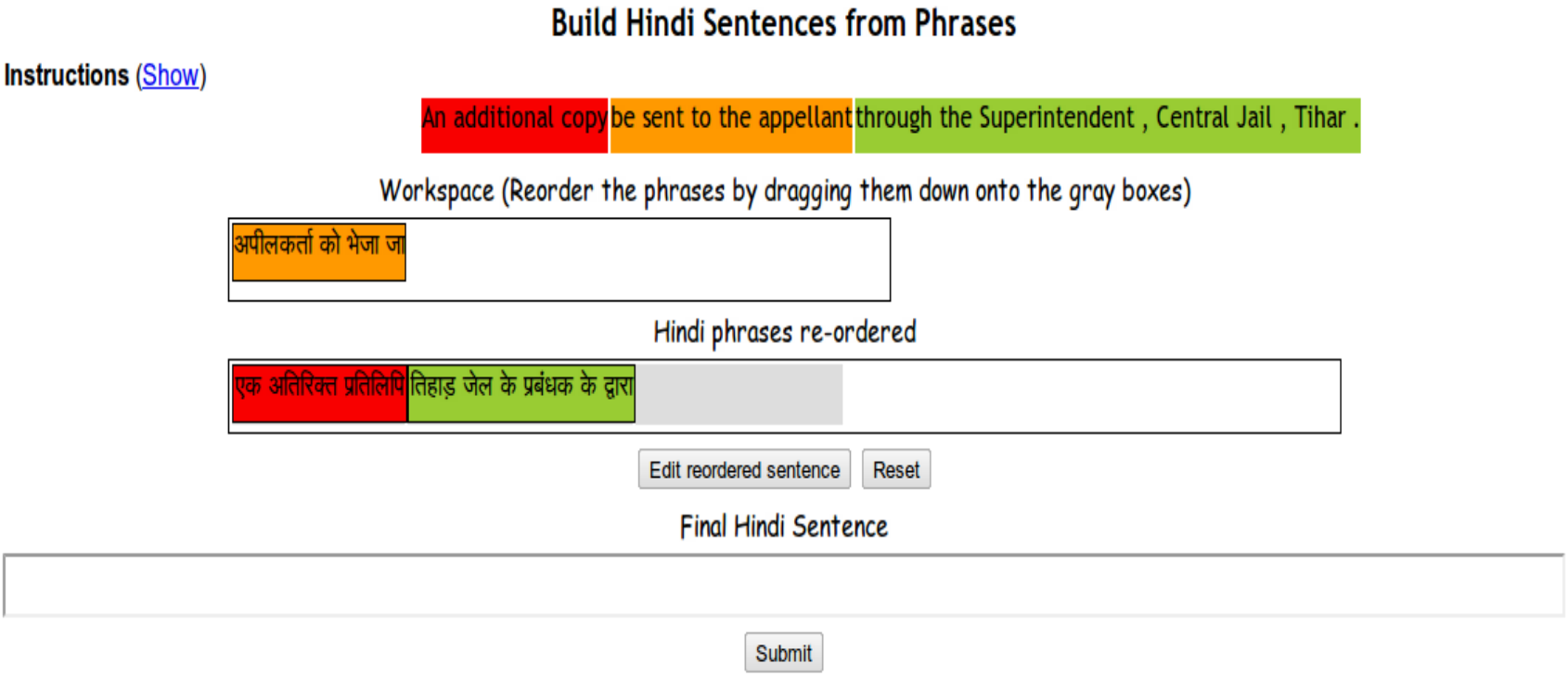
Phrase Validation

- In Phrase Validation, for a given source and target phrase the worker is required to assign a rating to the translation on a scale of 5 (0 means worst and 5 means best)



Phrase Re-combination

- In Phrase Re-combination, given a list of target phrases the worker is required to align the phrases and do necessary editing to obtain a valid target sentence.



Quality Control for Validation

- Spam Detector/ Noise Removal:
  - Each Validation Assignment will contain few spy phrases.
  - If a worker fails to successfully validate this spy phrases then the worker will be consider as spam and the Assignment will be rejected
  - The rejected Assignment will be resend to the crowd by the pipelining module.
  - After certain number of iteration if we are not getting valid assignment then respective measures will be taken.
  - If we get valid assignment then the rating will be used for analyzing quality of phrase translation.

Quality Control for Translation

- Based on the validation rating for same pair of source and target phrase an average score will be computed and assigned to that target phrase.
- The target Phrase which will have the highest score for a particular source phrase will be considered as the winner.
- If the winner target phrase if above a threshold value then it will be preserved else that source phrase will be resend to crowd for translation.
- After certain number of iteration if we are unable to get a good translation then respective measure will be taken.

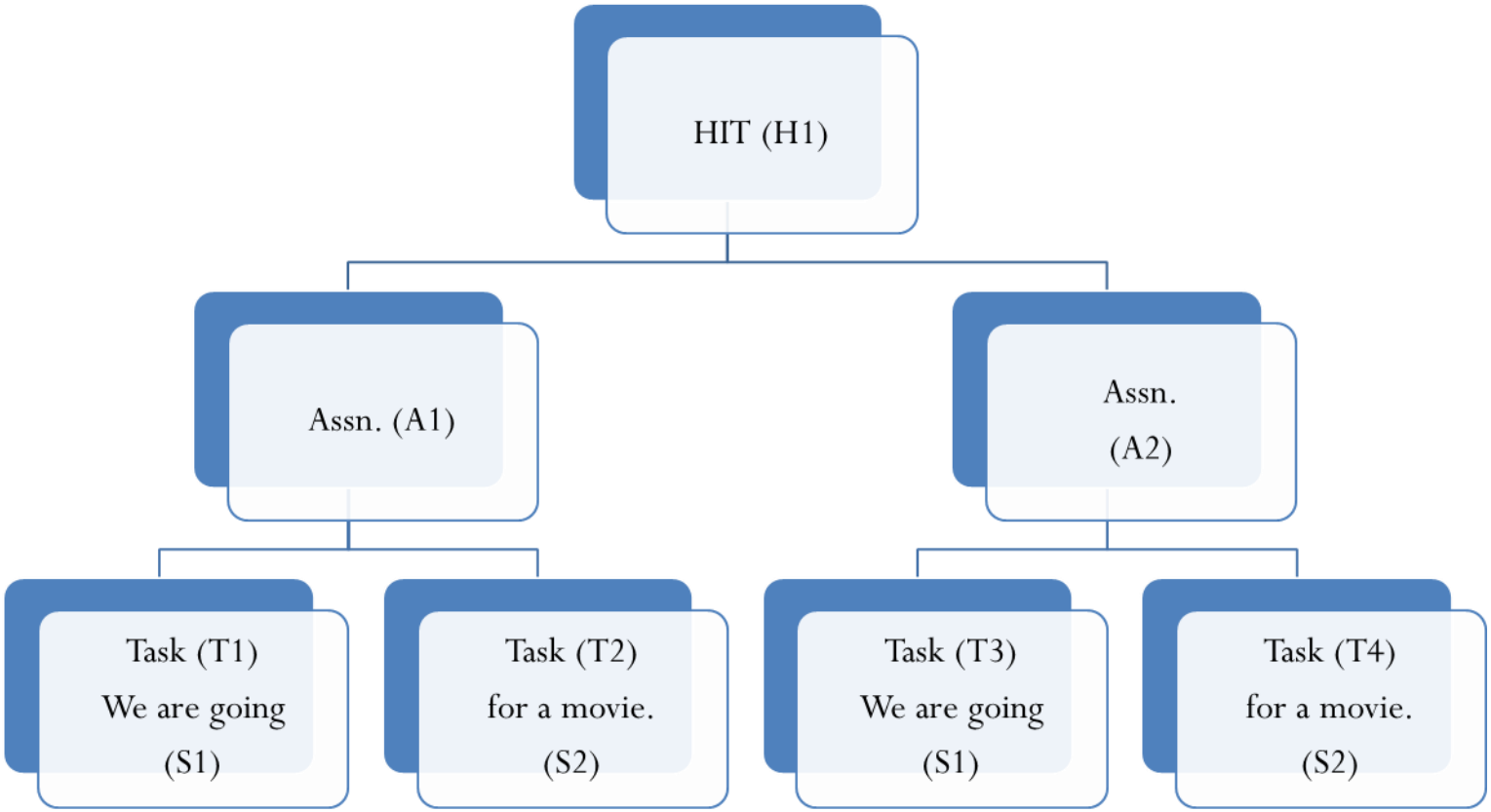
Quality Control for Re-combination

- Quality Control for Re-combination begins once the phrase re-combination phase is over.
- It checks for the proper alignment of phrases and whether the sentence is complete on its whole.
- Linguist may be introduced in this stage to get an authorized confirmation on the translated sentences.

Working example

Phrase Translation

- Sentence=We are going for a movie.
- S1=We are going, S2= for a movie.

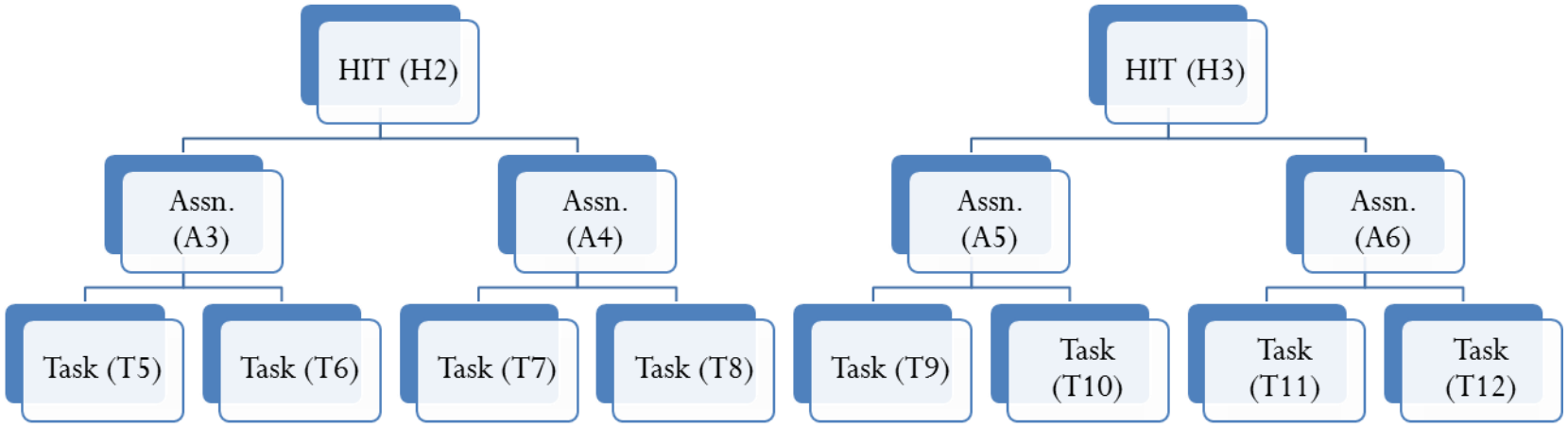


Tgt1= हम जा रहे है Tgt2= मूवी के लिए Tgt3= वे अरे गोइंग Tgt4= फॉर अ मूवी

Phrase Validation

S1=We are going S2=for a movie.

Tgt1= हम जा रहे है Tgt2= मूवी के लिए Tgt3= वे अरे गोइंग Tgt4= फॉर अ मूवी



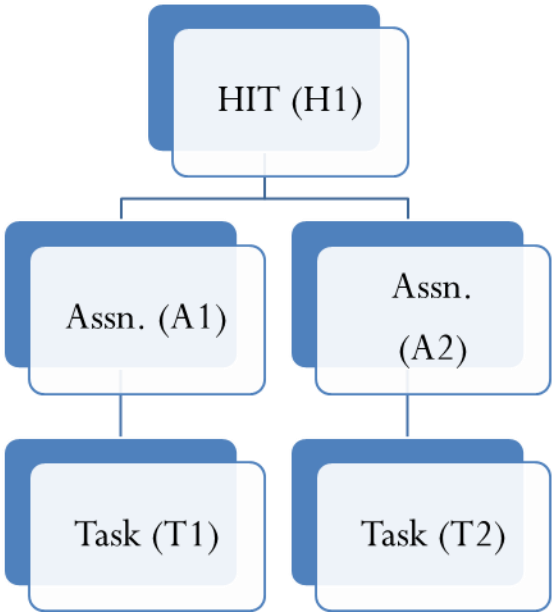
S1,Tgt1 R1=4 S2,Tgt2 R2=5 S1,Tgt1 R3=5 S2,Tgt2 R4=5 S1,Tgt3 R5=0 S2,Tgt4 R6=1 S1,Tgt3 R7=1 S2,Tgt4 R8=0

S1,Tgt1 = (R1+R3)=(4+5)/2=4.5
S1,Tgt3= 0.5
S2,Tgt2=5
S2,Tgt4=0.5

Phrase Re-combination

Source Sentence=We are going for a movie

Tgt1= हम जा रहे है Tgt2= मूवी के लिए



हम जा रहे है मूवी के लिए हम मूवी देखने जा रहे है

http://www.cse.iitb.ac.in/~abhijitmishra/reorder.html

Time for a Demonstration

Demonstration Steps:-