NLP Based Intelligence

Department of Computer Science and Engineering
(Supervisor: **Dr Deepti Gupta**)

'S

Name of Member	Roll Number	Role
Akash Kandpal	1513310027	NLP/ Deep Learning Part
Sivasish Koch	1513310214	UI / UX designer
Saif Ali	1513321165	Linking Pipelines and WebApp
Rahul Bilra	1513310161	Scraping the data

Index

- Introduction
- Literature Survey/Existing System
- Problem Statement
- Proposed Methodology
- Feasibility Study
- Hardware /Software Requirement
- Conclusion
- References

Introduction

- We are trying to make a platform where multiple services related to Human-Computer Interaction can be implemented and faster integration of these services is possible on your app, website or e-learning platform.
- Also, we'll be providing API's for different usecases with a limit per month.

Problems in Existing System

- Non-ability to detect handwritten notes and query on them.
- QA System implemented till now are too basic and simple.
- Automating tasks involving Vision tasks not much used in academics.
- Plagiarism detector mostly uses random algorithms like edit distance for docs and are not efficient.
- Chatbot integration is not that easy on your website.
- Resume Parsers are not that good and automation tasks for resume scanning are still not using ML.
- Translation of texts is limited specially when considering regional languages.

Problem Statement

Currently, we as a human want to get results faster and we have scarcity of time. So , when we have loads of repeated tasks we wish it could be automated and machines could understand our speech and our text in a better way.

Also ample of data while reading we want someone who could explain to us the important points or the whole idea of the complete document.

Consider a case, where we are given an unseen passage which is not even popular thus google won’t help. So where to find answers to questions which are not popular and one has data for that, one need to get someone else to go through the data and give them the data which is relevant to them as the user don’t have time to go throught the complete document.

In cases, where we have handwritten data, we don’t have softwares which can even detect it properly properly. So, loads of information which could have been beneficial cannot be automated, simlarly for Resume parsing tasks for HR is a headache.

Proposed Methodology

- The platform will be built on a Django Webapp with hosting the API’s on AWS Server.
- We’ll be building separate pipeline mechanisms(using PySpark) for most of the integration services.
- NLP part will be backed with separate corpus formation of about 10 gb disk space for QA System.
- Computer Vision part will be to done by implementing CNN and using libraries like tesseract from Python.
- Deep Learning will be used for Chatbot, where RNN will be used with LSTM approach.
- We’ll be providing following services :
- 1. MCQ’s Based Exams Checker
  2. Chatbots integration
  3. QA System and Text Summarization
  4. Text Translation and Transliteration
  5. Resume Parser and Scanner
  6. Checking Plagiarism

Feasability Study

- This project will be made on AWS Server but we’ll be using the free version of this, if we want to actually run into society then the cost will be there.
- It will take 8 -10 weeks to complete the project with 6-8 hours effort per day.
- Scalablity is pretty good as we will be using separate pipelines for each customer. And minimum API hits will be limited to 1 per 2 seconds.

Hardware/ Software Requirements

- Gensim, FASTTEXT and NLTK
- REST API’s : For hitting calls between ML model and app.
- Tensorflow : As a base for Keras and more optimization.
- Keras : For making Deep Learning Models
- Pandas : For cleaning the data
- Numpy : For mathematical purposes.
- Plotly : For visualising graphs.
- Scikit : For ML Algorithms.
- SHLDA : For Topic Modelling purposes.
- Open CV : For computer vision part.
- Convnet : For Handwritten Notes detection.

Conclusion

Our effort has been to make it easier for students, researchers and other people who have lots of data and don’t get time to read it. They can easily get relevant information from their data. This has never been approached before as previously people have tried to get the most relevant information from Google API’s and thus ignoring the individuality of every person.

Also, the approach involves further advancements like usage of Deep NLP and Computer Vision for detecting the words and getting relevant information out of it. Often a hybrid approach, judiciously blending apparently different techniques, provides improved results in the form of faster speed, increased relevancy, and higher precision and recall measures.

References

- Green BF, Wolf AK, Chomsky C, and Laughery K. Baseball: An automatic question answerer.
- Weizenbaum J. ELIZA - a computer program for the study of natural language communication between man and machine.
- Woods W. Progress in Natural Language Understanding - An Application to Lunar Geology.
- Bobrow DG, Kaplan RM, Kay M, Norman DA, Thompson H, and Winograd T. Gus, a frame-driven dialog system.
- Katz B. Annotating the World Wide Web using natural language.
- Clark P, Thompson J, and Porter B. A knowledge-based approach to question answering.
- Riloff E and Thelen M. A Rule-based Question Answering System for Reading Comprehension Tests.
- Reading Comprehension Tests as Evaluation for Computer-Based Language Understanding Systems, Vol. 6, 2000, pp. 13-19.
- Ittycheriah A, Franz M, Zhu WJ, Ratnaparkhi A and Mammone RJ. IBM’s statistical question answering system.

Thank You

Feedback from the Panel

- The panel wants the team to :