IASNLP-2011 Projects

Speech

Kishore Prahallad
kishore@iiit.ac.in

Speech

- Title: **Speaker identification on a large datasets**
- Team Size: 2
- Description: Speaker identification is a task of identifying a speaker from 1 out of N speakers using his/her voice samples. As the value of N becomes large, the complexity of identification task increases. In this project,. we attempt speaker identification on a large datasets using clustering techniques, which help to reduce the search space and decrease the complexity of identification task.

Speech

- Title: **Voice conversion**
- Team Size: 2
- Description: The objective of this project is to develop a voice conversion which morphs voice of one person to another.

Speech

- Title: **Speech summarization**
- Team Size: 2
- Description: The objective of this project is to develop unsupervised methods for speech summarization.

Linguistic Networks

Dr. Monojit Choudhury
monojitc@microsoft.com

Linguistic Networks

- Title: **Unsupervised discovery of function words in a language**
- Description: Words in a language are categorized into function and content words. The former, which is a closed class of words, play important grammatical function in a language, though they may not have a meaning of their own. Identification of function words is important for several applications including machine translation, information retrieval, parsing, OCR etc. In this project we will look at various corpus statistics, including network properties of the words in a language, to discover function words in a language through standard statistical and machine learning techniques.
- Team size: 4

Linguistic Networks

- Title: **Computational Study of Bollywood song lyrics**
- Description: Bollywood has been producing a few thousand songs every year for past six decades. We want to study this huge body of text, which is freely available on the web (the song corpus will be provided to the students), using corpus linguistic, network theory and language modelling techniques. We will use machine learning techniques to cluster the songs into different genre and build genre prediction models. We can also try to predict the author (lyricist) of the songs using supervised learning techniques. One of the interesting sub-goals of the project is to understand how the language of the Bollywood songs has been changing over past decades reflecting the underlying social changes.
- Team size: 4 (students can work in pairs on genre classification, lyricist prediction, and trend identification)

Parsing

Samar Husain
samar@research.iiit.ac.in
Prashanth Mannem
prashanth@research.iiit.ac.in

Parsing

Title: **Malt, MST and Bidirectional parsers**

Description: The projects involve building dependency parsers for Hindi, Telugu and Bangla using state of the art parsers like Malt parser, MST parser and bidirectional parser (BDP).

Team size: 6

Mentors:
Malt Parser - Prudhvi Kosaraju (prudhvi@students.iiit.ac.in)
MST Parser - Rahul Goutam (rahul_goutam@students.iiit.ac.in)
BDP - Prashanth Mannem (prashanth@research.iiit.ac.in)

Parsing

Title: **Iterative learning for POS tagging, chunking, NER and parsing**
Description: To try to mitigate the errors introduced in isolated NLP tasks using iterations to incorporate additional linguistic information during post initial iterations.

Team size: 4

Mentors: Manish Agarwal (manish_agarwal@students.iiit.ac.in)

Semantics

Dr. Soma Paul
soma@iiit.ac.in
Kiran Mayee
kiranmayee@research.iiit.ac.in

Semantics

Title : **Sanity checks for PurposeNet**

The project involves Data Sanity checks in the already populated PurposeNet ontology. The purpose Ontology has 2 parts : An Artifact Ontology and An action Ontology. The artifact ontology contains Artifact's descriptions and properties along with their
class relationships (Ex. SuperClass , SubClass etc.). This data needs to be
checked on the basis of various factors such as completeness , class relationships , referencing etc. Tasks involved are as follows :

Contd..

Semantics

Title : **Sanity checks for PurposeNet**

- Completeness Check :

A list of common and essential properties for each Artifact has been prepared and the data is populated with these. Many instances occur in which one or more properties are not applicable for a certain artifact. These fields need to be marked as NULL or Empty (as per the notation). This task would involve a completeness check on all the fields, i.e. checking whether all the fields have been marked (either respective value or NULL) and that no extra information (other than the specified properties) has been added to the ontology.

Example 1 : Entity : CAR doesn't have a property Viscosity. Thus that field is marked as NULL .

Semantics

Title : **Sanity checks for PurposeNet**

• Class Relationships :

Artifacts have other artifacts as components and accessories. There may be erroneous instances where 2 artifacts refer to each other in their components/accessories. This is a problem of circular referencing.

Example 3 : Entity : CAR has Entity : CAR_ENGINE as a component, if via some Human error , Entity : CAR_ENGINE also has Entity : CAR as it's component , then the data is corrupted. Thus one needs to check data consistency with the real world scenario, whether the information in the ontology is consistent with each other.

Semantics

Title: **Populating Action Ontology using Verbnet**

Description: The project involves extracting thematic roles for verb’s canonical forms from Verbnet and populating action ontology of PurposeNet with this information.

Team size: 2

Semantics

Title: **Mapping semantic roles to karaka relation using Verbnet**

Description: Each verb is associated with a set of theta roles in Verbnet. The task is to write heuristics for determining the karaka label for each noun with a particular theta role that are specified in the syntax frame of the verb in Verbnet.

Team size: 2

Morphology

Dr. Uma Maheshwar Rao
guraohyd@yahoo.com

Morphology

Title: Building a morphological analyzer

Description: This involves preliminary analysis of major word classes in a language and adapting an FST based morphological analyzer for your language.

Team size:4

Anaphora Resolution

Dr. Dipti Misra Sharma
dipti@iiit.ac.in
Praveen Dakwale
praveen.dakwale@research.iiit.ac.in
Himanshu Sharma
himanshu.sharma.rocky@gmail.com

Anaphora Resolution

Title: Anaphora resolution for Hindi

Description: Try out various methods to build an anaphora resolution system for Hindi

Team size: 4

Question Generation

Prashanth Mannem
prashanth@research.iiit.ac.in

Question Generation

Title: Question Generation from paragraph

Description: Given a paragraph, the task involves generating meaningful questions on the paragraph. T

he questions are evaluated along several dimensions including scope of the answer, grammaticality, semantic validity, question type correctness and diversity.

http://www.questiongeneration.org/QGSTEC2010

Team size: 2

Resource Creation

Dr. Dipti Misra Sharma
dipti@iiit.ac.in
Samar Husain
samar@research.iiit.ac.in
Rahul Agarwal
rahul_agarwal@students.iiit.ac.in

Resource creation

Title: Creating linguistic resource of different granularities

Description: Annotated form the bedrock of modern NLP/CL. Through this effort students will explore different levels of linguistic annotation such as POS tagging, morph, dependency etc.. and discover the theoretical and practical issues involved.

Team size:6

Transfer Based MT (TBMT)

Dr. Dipti Misra Sharma
dipti@iiit.ac.in
Piyush Agarwal
piyusharora@students.iiit.ac.in
Rashid Ahmad
rashid101b@gmail.com

TBMT

Title: Transfer based MT between Indian languages

Description: TBMT is one of the methods to doing machine translation where the overall task is divided into three broad steps (source analysis, transfer and target generation). In this project, the participants will work on specific issues one of these three steps.

Team size: 4

Statistical MT

Dr. Sriram Venkatapathy
sriram.venkatapathy@gmail.com
Prasanth Kolachina
prasanthk.ms09@gmail.com
Jayendra Rakesh
jayendra.rakesh@gmail.com

SMT

Title: Basic experiments in SMT

Description: SMT approaches have State of the art machine translation system. In this project, the participants will get an exposure to Moses (SMT testbed). Project involves basic experiments in handling the system and exploring various techniques supported by Moses.

Team size: 3

Sentiment Analysis

Prof. Sivaji Bandyopadhyay
sivaji_cse_ju@yahoo.com
Amitava Das
amitava.santu@gmail.com
Dipankar Das
dipankar.dipnil2005@gmail.com

Sentiment Analysis

Title: **Development of Emotion Lexicon and Word Level Emotion Analysis System**

Team size: 3

Title: **Development of SentiWordNet and Sentiment Classification**

Team size: 3

Preferred languages: Hindi, Telugu, Tamil and Urdu

Information Extraction, Information Retrieval (IE/IR)

Dr. Vasudeva Verma
vv@iiit.ac.in
Aditya Mogadala
aditya.mogadala@research.iiit.ac.in

IE/IR

Title: **Amazon data (search customer reviews)**

Description: Given the amazon.com data containing reviews, provide the search capability for finding the users who are relevant to the query provider.
For ex: "School of Hillel" should list all the users who have written about it. Assume query contains max. 3 words.

This allows us to find all the users who have common views on a product.

Team size: 2

IE/IR

Title: **Twitter data clustering based on categories**

Description: Twitter data consists of tweets posted by various users. The task is to cluster the tweets into different categories and identify the topic or a particular event that the tweet talks about.

Team size:2