CS6320.003 Natural Language Processing

Linguistic Learners

Project short Description

Ramkumar Paranthaman rxp152630
Vadivel Selvaraj vxs154530

Problem Description

Identify future calendar events with date and time from text. Five events – Marriage, Birthday Party, Meeting Anniversary, Seminar will be included in scope.

Proposed Solution and Implementation Details

1) Baseline system

Using Bag of words approach, pre-defined set of Keywords will be matched across text and based on presence of certain keywords, events will be identified.

2) Improvement Strategy

a. Lexical features

- Tokenizer
- Spell correction

b. Syntactic Features

- POS - Temporal expression tagging (sequence, duration and range) & POS tagging
- Syntactic pattern - Look for past tense tag (eg., VBD, VBN, etc.,) and ignore them as they are past events

c. Semantic Features

- Synonymy – from NLTK WordNet – to retrieve words that are synonymous to required events (Marriage, Birthday, Meeting, Anniversary, Seminar)
- Named Entity recognition to find location of the event

3) Examples

*a) **Naive approach –** The lecture starts at 11.00 A.M in Auditoruim*
output: Not an event
output is wrong because none of the words match with the keywords [start, end, meeting]

*b) **Lexical Features –** The lecture starts at 11.00 A.M in Auditoruim*
output: [*'The', 'lecture', 'starts', 'at', '11.00', 'A.M', 'in', 'Auditorium'*]

c) ***Syntactic Features*** *-The lecture starts at 11.00 A.M in Auditorium*
output: [('The', 'DT'), ('lecture', 'NN'), ('starts', 'VBZ'), ('at', 'IN'), ('11.00', 'CD'), ('A.M', 'NNP'), ('in', 'IN'), ('Auditorium', 'NNP')]
('11.00 A.M', <*Time*>) - result of temporal expression tagging (tag name depends on the tool)

*d) **Semantic Features**- The <u>lecture</u> starts at 11.00 A.M in Auditorium*
output: The <u>Seminar</u> starts at 11.00 A.M in Auditorium

4) Programming Tools

NLTK – WordNet, Spell correction, Timex, Named Entity Recognition

5) Architecture Diagram