NLP Project Summary Report

Team Name: Mysterious Meerkat
Team Members: Debjyoti Paul

December 11, 2015

System Components:

Debjyoti Paul, Mysterious Meerkat

Directory Structure and roles of each file

Note: This is a single member group, all the contribution is made by me. Hence skipping member contribution part.

- **qa.py** Question Answer System Main Driver
- **modules**
  - AnswerUtility.py Answer Formulation Module
  - SimilarityModule.py Sentence Score Aggregator Module
  - synonym_similarity.py Sentence Similarity Module based on Synset, Hyperset
  - ner_similarity.py NER and Rule based Similarity Module
  - QuestionClassifier.py Question Processing Main Module
  - MVectorizer.py Utility for Question Module
  - scikit_classifier.py Utility for Question Module
  - keywords_pickle_builder.py Utility for Question Module
  - coref.py BART webservice and passage preprocessor
- **papers**
  - QuestionClassifier.pdf Question Classifier Paper
  - SentenceSimilarity.pdf Sentence Similarity Paper
  - bart_coref.py Bart coref paper
- **poster**
  - qa_poster.pdf QA System poster
- **prepare.sh** Auto dependency installer
- **qcdata** Question Classifier training and test data

External resources:

Python Packages : *nltk, numpy, scipy, scikit-learn, pickle, requests, beautifulSoup4*
Coreference : *BART Beautiful Anaphora Resolution Toolkit*
Papers followed:

- *Li, Xin, and Dan Roth. "Learning question classifiers: the role of semantic information." Natural Language Engineering 12.03 (2006): 229-249.*
- *Li, Yuhua, et al. "Sentence similarity based on semantic nets and corpus statistics." Knowledge and Data Engineering, IEEE Transactions on 18.8 (2006): 1138-1150.*

Afterthoughts:

Success

1. Question classifier to determine answer type as HUM:ind, HUM:gr, LOC:city, NUM:date, NUM:money etc. is the key to enhance performance. Please refer to Figure 1 for categories.
2. Using regex to find number, date and currency such as 'dd[st|nd|rd|th] [Month]' in date, 1800-2100 for years and $100.0 or $ 0.50 for currency has helped.
3. The idea of measuring the length of the shortest path in the semantic ontology between two words has been beneficial. Also, extracting verb from question and finding matching verbs or its synset has improved system performance. Please refer to Table 1 for detail in Appendix

Regrets

1. Coreference resolution has little effect on my QA system, more manpower and time might have been helpful to test scoring functions thoroughly and revamp it.

Appendix

Note: Extremely sorry for adding more pages but thought this Appendix will be important

Class	#	Class	#
ABBREV.	9	description	7
abb	1	manner	2
exp	8	reason	6
ENTITY	94	**HUMAN**	65
animal	16	group	6
body	2	individual	55
color	10	title	1
creative	0	description	3
currency	6	**LOCATION**	81
dis.med.	2	city	18
event	2	country	3
food	4	mountain	3
instrument	1	other	50
lang	2	state	7
letter	0	**NUMERIC**	113
other	12	code	0
plant	5	count	9
product	4	date	47
religion	0	distance	16
sport	1	money	3
substance	15	order	0
symbol	0	other	12
technique	1	period	8
term	7	percent	3
vehicle	4	speed	6
word	0	temp	5
DESCRIPTION	138	size	0
definition	123	weight	4

Figure 1: Question Classifier classes; Course class in bold followed by fine classes hierarchy

Table 1: Synset Mapper

class	synsets
ENTY:animal	noun.animal, noun.person
ENTY:body	noun.body, noun.person
ENTY:color	noun.color
ENTY:cremat	noun.congnition, creation.n.02 writing.n.02, publication.n.01, music.n.01
ENTY:currency	noun.currency, noun.quantity
ENTY:dismed	noun.illness, noun.medicine, noun.drug, noun.health
ENTY:event	noun.event, noun.holiday
ENTY:food	noun.food
ENTY:instru	noun.instrument, noun.artifact, noun.object
ENTY:lang	noun.language, noun.communication
ENTY:letter	noun.communication, noun.letter
ENTY:other	noun.other
ENTY:plant	noun.plant
ENTY:product	noun.product, noun.object, noun.artifact
ENTY:religion	noun.religion, noun.group
ENTY:sport	noun.sport, noun.game
ENTY:substance	noun.element, noun.object, noun.substance
ENTY:symbol	noun.symbol, noun.sign
ENTY:techmeth	noun.technique, noun.method
ENTY:veh	noun.vehicle
DESC:desc	noun.motive
DESC:manner	noun.manner
HUM:gr	noun.group, noun.organization
HUM:ind	noun.person
HUM:title	noun.person
HUM:desc	noun.person
LOC:city	noun.location, noun.city
LOC:country	noun.location, noun.country
LOC:mount	noun.location, noun.mountain
LOC:other	noun.location
LOC:state	noun.location
NUM:code	noun.phone_number, noun.code
NUM:date	noun.time_period, noun.date, noun.time
NUM:money	noun.monetary_unit, noun.currency, noun.money
NUM:ord	noun.chapter, noun.rank
NUM:other	noun.number
NUM:period	noun.time_period, noun.time_unit, noun.time
NUM:perc	noun.percent
NUM:speed	noun.speed, noun.rate
NUM:temp	Fahrenheit.a.01, Celsius.n.01
NUM:size	linear_measure.n.01
NUM:weight	mass_unit.n.01