Author: Ishu Dharmendra Garg Mentors: Salil R Joshi, Vikas V Joshi Manager: Sameep Mehta

Motivation

- Companies like IBM for the task of Information Retrieval and Sentiment Analysis usually write Rule Annotators for mining the information in the form of structured data
- Most often these rules are manually written by someone which require a lot of human expertise over the domain
- Writing such annotators requires a lot of time and resources
- The annotators which are written are generally applicable to a particular domain and for a different domain a whole new set of annotators have to written

Definitions

- Dictionary: A set of tokens (words/phrases). For example: Organization {IBM, Box, Microsoft, …}
- Template: A sequence of dictionaries. For example: <Organization> <Hire> <Person>
- Concept: A structured representation of an entity type or relation that can be collectively captured using one or more templates

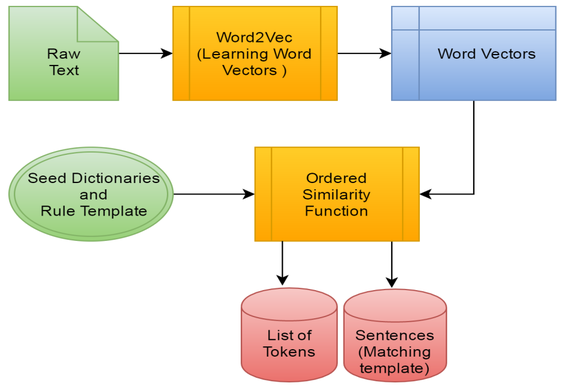
Problem Statement

Simply speaking we want to automate the annotation task. (Minimize human effort and increase cross domain learning)

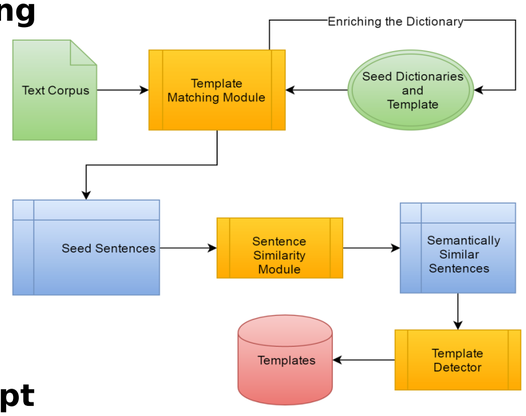
- Template Matching: Given a template, mine all the sentences from the corpus that follow the template
- Concept Learning: Given a template, find some other templates that can represent the same concept. For example, given a template for mining 'acquisition' event, find other syntactical constructs and corresponding templates
- Concept Abstraction: Given a template representing some concept in one domain, find semantically / syntactically generalized representations in same or different domain

Architectures

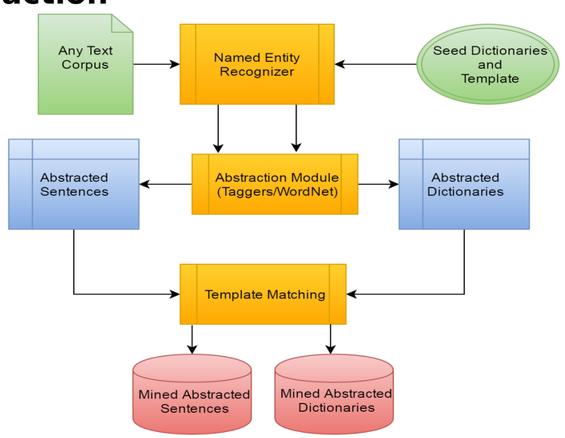
Template Matching



Concept Learning



Concept Abstraction



Novelty

- Independent of language, domain or source of domain data.
- Semi-Supervised but unlabelled Learning.
- Context of the words determined from the domain data and also used in template matching and mining which makes it different from naive implementations.
- Word2Vec uses shallow neural networks making the process computationally efficient and scalable.
- Works with very little human input and supervision.

Things that did not work

- Several Word vectors tried but only word2vec worked.
- State of Art Sematic Similarity tools are not good enough
- Different types of ordered similarity functions tried but only a few worked.

Results

We were successfully able to develop a ML pipeline to mine sentences, events and entities from raw unstructured text in any natural language achieving precision and recall of 0.76 and 0.65 on AP news Corpus.

Future Work

- Improving Recall and Precision: Though the recall and precission are satisfactory, but further improvements can be made.
- Template Generation: Find a way to generate diverse templates similar to a given tamplate so that user do not have to be creative to generate templates.