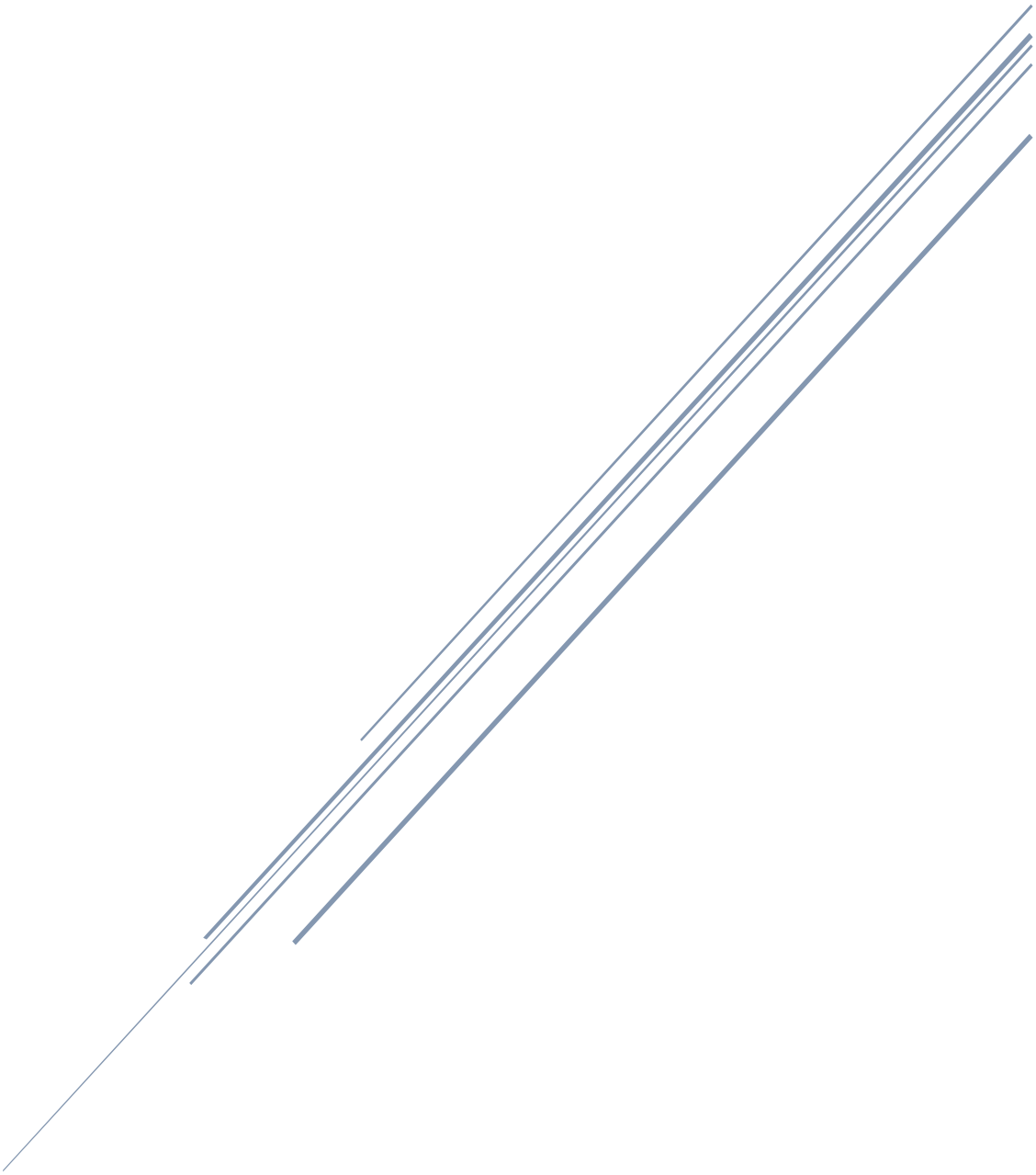
PURPOSE EXTRACTION TOOL

Prof. Soma Paul



Aditya Sanka 201001135
N.Mithun Kumar 201001072

About

Every artifact is created to serve a specific purpose which can be called as its primary purpose. The project aims at extracting primary purpose of artifacts from web corpus. An artifact can be used for multiple purposes but it has a primary purpose. This primary purpose can also be called as the utility of the artifact. Significant work has been done on semantic analysis in the field of Natural Language Processing. This paper aims at extraction of purpose data from web.

Prior Work

Eugene Agichtein , Luis Gravano have developed the Snowball system which automatically generates patterns and extracts tuples from large collections of text data. Kiran Mayee, Rajeev Sangal and Soma Paul used the concept of surface text patterns for extracting purpose data from a web corpus.

Introduction

We have implemented a system that extracts purpose information from web corpus using surface text patterns. The system aims at extracting purpose information from sentences in which the purpose is explicitly mentioned. For example "Pen is device for writing on paper" is explicit whereas "Airbase is a base for military aircraft" is implicit in nature.

We studied the data for patterns that indicate the presence of purpose information. These patterns could be effectively represented using various methods. Regular expressions are effective in pattern matching with strings. These regular expressions can be further strengthened by the addition of the respective POS tags to the expression. This combination of Regular Expression and POS tags is called a surface text pattern. These patterns are matched against the data extracted for a given artifact and its purpose is identified.

Our approach:

The extraction of primary purpose of an artifact involves two stages:

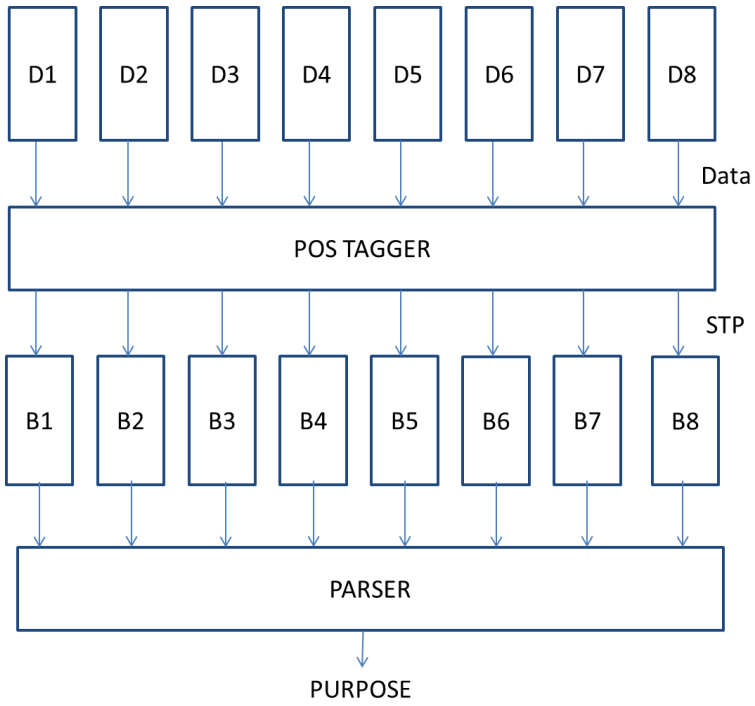
1. Information Retrieval
2. Purpose extraction

Information Retrieval:

When an artifact is given as input, the definitions are extracted from selected web dictionaries. We observed that most of the definitions for an artifact contained its purpose information. Several artifacts were given to a generic search engine and the first thousand results were stored. Most of these results did not contain any purpose information. Even when found either the information was implicit in most of the cases. We observed that when the information was explicit, it was limited to few sentences. We selected domain independent online dictionaries so that the system will work for all domains.

Web Scraping techniques were used and scripts were developed for each of our selected online resources. These scripts are independent of each other and run in a parallel environment. These results are used in later stages of the system for purpose extraction.

Previous work using surface text patterns involved downloading the first thousand documents from the result of a search engine. This would not be efficient as most of documents do not have the required information. Our approach of using selected resources decreases the time involved in information retrieval by a large extent. Since the result is a list of definitions from dictionaries the results are also reliable.



D1 to D8 : Online dictionaries.

B1 to B8 : Surface text patterns.

Purpose Extraction:

A surface text pattern is a combination of Regular expression and POS tags. We initially formed a training corpus of sixty artifacts. This corpus was studied for the presence of information and patterns were manually formed. These patterns are represented using the concept of surface text patterns. A testing corpus of fifty artifacts was formed. Each of the patterns are matched against the entire testing corpus and results were used to find the coverage and accuracy of that respective pattern. Coverage is the percentage of artifacts for which we have result. Accuracy is the percentage of correct results in total for a given pattern. Similar patterns were grouped together and sorted according to their accuracies.

For example when

“Toothpaste is a paste for cleaning teeth“

is matched against

“a DT \w*? NN for IN \w*? NN \w*? NN”

The result would be “cleaning teeth”.

The extracted text data is tagged using a POS tagger. Each sentence is now matched against all groups of patterns. Once a match is found in a group we stop and move on to the next sentence in the corpus. We can observe that processing of each of the sentences is independent of other. So all these sentences are parsed in a parallel environment. This decreased the execution time of the program by a large extent.

We have also implemented a cache using simple caching techniques. This cache would consist of the final results of all the pre-processed artifacts. When we encounter an artifact that is already in the cache we just retrieve the result from cache. This drastically improves our performance as we process the artifact only for the first time.

Since both the stages of the system run in a parallel environment, the entire program is parallel and has an increase in performance with respect to time.

Results of purpose extraction module:

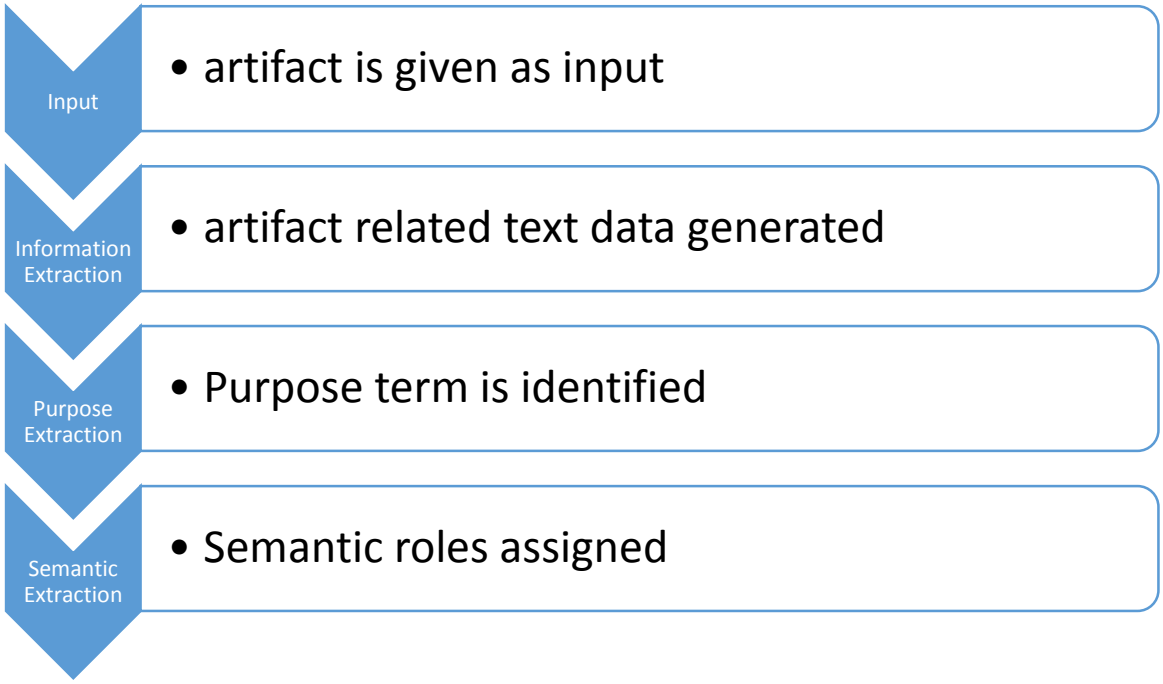
Our entire corpus consists of a total of 110 artifacts. We divided it into training data of fifty artifacts and testing data of sixty artifacts. The results when our system was run on test data were :

Coverage : 49/60 82%

Accuracy : 45/49 92%

Purpose extraction tool can be divided into three major modules :

- Information Extraction
- Purpose Extraction
- Semantic Extraction



Extraction of semantic roles:

This module deals with identifying the participants involved and assigning proper semantic roles to them. As we know the nature of the information we are dealing with, we can expect the semantic information present in the text data. We again use the concept of surface text patterns for the extraction of semantic roles. The following semantic roles are to be extracted

Agent

Theme

Instrument

Location

Any other semantic information available in the corpus is also extracted.

We will be using Verbnet for the extraction of semantic roles. Verbnet is the largest online verb lexicon available for English. Verbnet is domain independent and hence the generality of the tool developed will not be compromised. Verbnet has syntactic and semantic information regarding the listed verbs. Verbnet comprises of a total of 5800 verbs grouped into 275 classes. Each verb can be a member of multiple classes. Each class contains several frames which have semantic as well as syntactic information. The hierarchy of Verbnet can simply be depicted at three levels

- Verbnet
    - Class
        - Frame

Each frame can be identified using a sequence of POS tags. We tag the corpus using the standard nltk parser and we get the syntactic structure of the available text data. We match the frames against our information and extract the semantic roles. We observed that a verb cannot have the same syntactic structure in different classes. We can conclude that the output of the module is unique for a given sentence .Taking these observations into account we designed an algorithm for semantic role extraction.

Algorithm :

- Purpose term is given as input.
- Identify the classes to which the verb belongs.
- List the frames available in these classes.
- Tag the text data according to the syntactic structure of Verbnet.
- Match each sentence in tagged data against each frame.
- Once a match is found extract the semantic roles.

Example :

"Smith hit the ball with a bat."

When "hit" is given as input to the algorithm we find that "hit" belongs to the class hit-18.1 in the Verbnet corpus. We then match the sentence against the frame with "NP V NP PP.instrument" as the primary structure and secondary structure "NP-PP; Instrument-PP" .The semantic roles extracted are:

Agent : Smith

Patient : ball

Instrument : bat

Future Work :

Implementation of the semantic extraction module.

Generalisation of the tool.