# Manual Rule-based Extractor Baseline

Following combinations of POS and word based regular expression were used.

* *attended/VBD the/DT*
* *was/VBD a/DT student/NN at/IN*
* *was/vbd educated/vbn at/in*
* *was/vbd student/vbn at/in*
* *attended/VBD the/DT*
* *where he/PRP attended/VBD*

Regexes over word

* “(.\*?) educate.\*”
* “(.\*?) graduat.\*”
* “(.\*?) complet.\*”
* “(.\*?) attend.\*”
* “(.\*?) student.\*”

Regexes over POS

* “(.\*?) VBD PRP.\*”
* “(.\*?) VBD DT.\*'”
* “(.\*?) VBD VBN.\*”
* “(.\*?) VBD.\*IN.\*”
* “(.\*?) PRP.\*VBD.\*”

Overall syntactic structure of the sentence is ignored by the regex based pattern matching and thus solely relies on the regex words/POS tags under consideration, Positive by regex doesn’t consider the subject, object in question.

“ Koji Nakano (b. August, 1974) is a Japanese composer. He was born in Japan and educated in Boston, The Hague, and San Diego. Mr. Nakano has been recognized as one of the major voices among Asian composers of his generation. His work strives to merge Western and Eastern musical traditions, and reflects the relationship between beauty, form and imperfection through the formality of music. Mr. Nakano received his Bachelor's Degree in composition with distinction, and Master's Degree in composition with academic honors and distinction, Pi Kappa Lambda, from the New England Conservatory of Music in Boston, where he studied with Lee Hyla and John Harbison. Later, he studied with Dutch composer Louis Andriessen in Amsterdam and at the Royal Conservatory of Hague as the Japanese Government Overseas Study Program Artist. Mr. Nakano received his Ph.D. in composition from the University of California at San Diego, where he studied with Chinary Ung. In addition to being the recipient of The American Artists and Museum Professionals in Asia Fellowship from the Asian Cultural Council, Mr. Nakano is also The first recipient of the Toru Takemitsu Award in Composition from the Japan Society of Boston awarded annually to the most talented young composer in the Boston area. In 2008, he became the first composer to receive the S&R Washington Award Grand Prize from the S&R Foundation, which is awarded annually to the most talented young artist (in the fields of fine arts, music, drama, dance, photography and film), for his/her contributions to U.S.- Japanese relations. The past distinguished grand prize awardees include soprano Maki Mori (2000), pianist Yu Kosuge (2002), violinists Yosuke Kawasaki (2004), Sayaka Shoji (2006), and Tamaki Kawakubo (2007).”

Regex model predicts this a ‘yes’ for this case since its find studied at’ as well as POS tag sequence “VBD IN” in the underlined intermediate text part and hence gives as erroneous prediction.

|  |  |  |  |
| --- | --- | --- | --- |
| Experiments | F-Score | Precision | Recall |
| Word | 0.782 | 0.763 | 0.812 |
| POS | 0.802 | 0.792 | 0.843 |
| POS+WORD | 0.791 | 0.782 | 0.824 |

# Brown Clustering:

*Brown clustering* is a method used to create clusters of words that are similar. It is an instance of a clustering algorithm which generates a hierarchical cluster of words. Every word is assigned a bit-string. The bits to the left are the most significant. If a word's significant bits agree with another word's, those words will be in a similar category.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Experiment(Prefix) | Clusters | Precision | Recall | F-Score |
| 4 | 12 | 0.670 | 0.870 | 0.785 |
| 6 | 49 | 0.682 | 0.910 | 0.791 |
| 8 | 49 | 0.710 | 0.801 | 0.791 |

Issues:

I have tried to use Brown clustering to group syntactically similar words together. While the clustering is highly dependent on the merging criterion, the resulting clustering is usually not syntactically similar words. Hence it fails to capture syntactical insights. Decreasing prefix length forms more generic clusters hence the drop in F1- Score.

# Bag of words:

Bag-of-words Baseline

Use the training feature vectors to train an SVM classifier (see Classifier section below for details) and make predictions on the test. Evaluate the performance of the classifier with these BOW features

Code provided develops features that has count of occurrence of a particular token in intermediate text.

Additionally I have used word tokenizer and tree bank tokenizer for split function. This improved performance and results because tokenization was performed based on prior learning.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Experiments | C | Precision | Recall | F-score |
| Plain BOW | 1 | 0.773 | 0.791 | 0.773 |
|  | 0.01 | 0.753 | 0.721 | 0.751 |
|  | 10 | 0.741 | 0.832 | 0.778 |
|  | 100 | 0.738 | 0.791 | 0.765 |
| NLTK Tokenize | 1 | 0.778 | 0.795 | 0.783 |
|  | 0.1 | 0.771 | 0.782 | 0.792 |
|  | 10 | 0.769 | 0.811 | 0.795 |
|  | 100 | 0.752 | 0.799 | 0.789 |

Issue and Errors

This approach fails to capture any syntactic features or dependency based features. High Frequency words will dominate the result.

Eg:

Salendine Nook High School Siddique At the age of 11 years, Siddique studied Salendine Nook High School, a multicultural school, where she excelled in English. She later moved to Greenhead College. studied

Here ideally this sentence doesn’t contain a relation but the model says it has a relation, which is wrong. This is because it has high frequency words like studied.

# Dependency Parse:

Implemented Using Stanford CoreNLP.

Steps:

1. Form a graph from dependencies returned by Stanford parser.
2. Find the shortest path between Person and Institute.
3. Report POS Tags, dependency tags, shortest Path Length, NSubjPass dependency tag.

Eg: Gerald Flurry graduated from Ambassador College, Pasadena, California, in 1970 and became a minister with the Worldwide Church of God (WCG) in 1973. In 1975 he was transferred to Pasco, Washington. Eventually he transferred to Oklahoma in 1985. During the three years after Herbert Armstrong's death in 1986, WCG made several doctrinal changes that Flurry objected to as doctrinally false. He began to make known his opposition to these changes and produced a manuscript that would become the book, Malachi's Message to God's Church Today. These events led to his being summoned by WCG leaders to appear before them, where Flurry was fired from the WCG on December 7, 1989. From 1992 onwards he has taught that this booklet is the 'little book' of Revelation 10.

Shortest path: [1, 2, 3, 6]  
Dependency Tags for above Path : ['compound', 'nsubj', 'nmod:from']  
Node 1    token\_text: Gerald      pos:NNP  
Node 2    token\_text: Flurry      pos:NNP  
Node 3    token\_text: graduated  pos:VBD  
Node 6    token\_text: College      pos:NNP  
Path Length : 4

Verb Count : 1

Subject Dependency : True.

Features Used:

1. Path Length: Formed a graph from dependency edges and used source as the person’s name and destination as university name. Then find the shortest path between source and destination. Use this length of path as a feature.
2. Dependency Tags: Form a feature like bag of words by forming a feature vector of size equal to length of available dependencies. Then increment the count of individual dependencies found in the shortest path.
3. Verb Count: Shortest Path Can contain more the one verb, so using this information.
4. Subject Dependency: Checking if the person occurs in any subject based dependency in the document.

Test Results:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Experiment | C -Value | F-score | Precision | Recall |
| SVM(Linear function) | 0.1 | 0.817 | 0.792 | 0.782 |
| SVM(Linear function) | 10 | 0.803 | 0.794 | 0.791 |
| SVM(Linear function) | 1 | 0.805 | 0.782 | 0.799 |
| SVM(Linear function) | 0.01 | 0.791 | 0.782 | 0.789 |

Insights:

Dependency features helps capture syntactic information which help the model perform better.

Maribyrnong College   Scud He has had a minor career in modelling and starred in the American reality television dating show Age of Love. He is nicknamed 'the Scud', after the Scud missile. He was educated at Maribyrnong College.', after the Scud missile. He was educated at

Model fails to tag this as negative example

Kitchen Sink:

I have tried various combinations of Kitchen sink experiments. Experiments performed as follows:

Bag of Words(NLTK Tokenize) + Dependency Features + Brown Clustering:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Experiments | C value | Precision | Recall | F-Score |
| SVM Linear Function | 10 | 0.794 | 0.821 | 0.809 |
| SVM(Radial Function | 10 | 0.797 | 0.811 | 0.803 |
| SVM Radial Function | 0.1 | 0.744 | 0.771 | 0.798 |
| SVM Linear Function | 0.1 | 0.794 | 0.821 | 0.809 |

Bag of words + Dependency:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Experiments | C value | Precision | Recall | F-Score |
| SVM Linear Function | 10 | 0.734 | 0.765 | 0.778 |
| SVM (Radial Function | 10 | 0.752 | 0.798 | 0.771 |
| SVM Linear Function | 0.1 | 0.789 | 0.791 | 0.769 |
| SVM Linear Function | 1 | 0.764 | 0.781 | 0.782 |

Bag of words + Clustering

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Experiments | C value | Precision | Recall | F-Score |
| SVM Linear Function | 10 | 0.754 | 0.773 | 0.763 |
| SVM(Radial Function | 10 | 0.743 | 0.781 | 0.759 |
| SVM Linear Function | 0.1 | 0.761 | 0.786 | 0.761 |
| SVM Linear Function | 1 | 0.771 | 0.772 | 0.769 |

Dependency Features + Clustering:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Experiments | C value | Precision | Recall | F-Score |
| SVM Linear Function | 10 | 0.792 | 0.824 | 0.834 |
| SVM(Radial Function | 10 | 0.797 | 0.811 | 0.829 |
| SVM Linear Function | 0.1 | 0.782 | 0.792 | 0.821 |
| SVM Linear Function | 1 | 0.794 | 0.821 | 0.848 |

Above experiments show that the best configuration for kitchen sink is combination of brown clustering and Dependency features. One reason could be combination of frequency property added by brown clustering but making sure that the model doesn’t have a bias for a word. Additionally, dependency features help capture syntax based features.

Other combinations show moderate results.

Worst performance is provided by bag of words and brown clustering configuration since it fails to capture the syntactic information. Therefore, the model performs poorly.

|  |  |  |  |
| --- | --- | --- | --- |
| Experiment | F-Score | Precision | Recall |
| Manual RegExp | 0.791 | 0.782 | 0.824 |
| Bag of words | 0.792 | 0.771 | 0.782 |
| Brown Clustering | 0.791 | 0.710 | 0.801 |
| Dependency Features | 0.817 | 0.792 | 0.782 |
| Kitchen Sink | 0.848 | 0.794 | 0.821 |