ABSTRACTIVE MULTI-DOCUMENT TEXT SUMMARIZATION

TEAM MEMBERS :-

AKSHATA BHAT - 1PI11IS008

K R ANUSHREE - 1PI11IS127

GUIDE

DR. S NATARAJAN

PROFESSOR

Dept. of ISE, PESIT

Bengaluru

Contents

Introduction

- ***Text Summarization*** is the process of extracting salient information from the source text and to present that information to the user in the form of summary.

- ***Multi-Document Summarization*****,** a form of Text Summarization, is an automatic procedure aimed at extraction of information from large cluster of documents about the same topic and generating a concise summary.

Introduction

- ***Extractive Summarization*** method consists of selecting important sentences, paragraphs etc. from the original document and concatenating them into shorter form. The importance of sentences is decided based on statistical and linguistic features of sentences.

- ***Abstractive Summarization*** consists of understanding the source text by using linguistic method to interpret and examine the text. The abstractive summarization aims to produce a generalized summary, conveying in information in a concise way.

Problem Definition

- To analyze the performance of some of the existing techniques for document summarization, by executing.
- To experiment with new machine learning techniques, which have not been used for summarization.
- To design a technique to generate Abstractive Multi-Document Summarization from the Extractive summaries generated.
- Evaluate their performance on the basis of few pre-defined parameters. This evaluation will be done manually.

Motivation

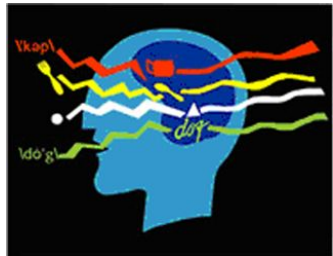
- Digital data available to us on World Wide Web is growing at an exponential pace.
- Text summarization across several documents which cover same subject has become an important and timely tool for user to quickly understand the large volume of information.
- For humans, generating a summary is a straightforward process but it is time consuming.
- There is need for automatically generating the summary.

Domain - Machine Learning and Natural Language Processing

- Machine learning is a type of artificial intelligence (AI) that provides computers with the ability to learn.
- Machine learning focuses on the development of computer programs that can teach themselves to grow and change when exposed to new data.
- Broadly divided into 2 categories :
  - Supervised learning
  - Unsupervised learning
- Between supervised and unsupervised is semi-supervised learning

Domain - Machine Learning and Natural Language Processing

- Natural language processing (NLP)- a component of artificial intelligence, is an ability of a computer program to understand human speech as it is spoken.
- Development of NLP applications is challenging- not always precise and often ambiguous.
- Major tasks - summarization, NER, Machine translation, Natural Language Generation (NLG), etc.
- NLG - a system that converts a computer based representation into a natural language representation.
- Current approaches to NLP are based on machine learning.



Literature Survey

TITLE	AUTHOR	PUBLICATIONS	DETAILS
A REVIEW ON ABSTRACTIVE SUMMARIZATION METHODS	ATIF KHAN, NAOMIE SALIM	JOURNAL OF THEORETICAL AND APPLIED INFORMATION TECHNOLOGY. (2014)	DISCUSSES THE STRUCTURE BASED AND SEMANTIC BASED METHODS AND THE STRENGTHS AND WEAKNESSES OF EACH METHOD. IDENTIFIES THE OPEN RESEARCH ISSUES.
AUTOMATIC TEXT SUMMARIZATION: PAST, PRESENT AND FUTURE.	HORACIO SAGGION AND THIERRY POIBEAU	THEORY AND APPLICATIONS OF NATURAL LANGUAGE PROCESSING, SPRINGER BERLIN HEIDELBERG (SOURCE : SPRINGER) . (2013)	DISCUSSES THE DIFFERENT APPROACHES DEVELOPED IN PAST. DISCUSSES THE MAJOR EVALUATION TECHNIQUES USED. IDENTIFIES THE ACTIVE RESEARCH TRENDS.
THE ELEMENTS OF AUTOMATIC SUMMARIZATION	DANIEL JACOB GILLICK	THESIS SUBMITTED TO COMPUTER SCIENCE, UNIVERSITY OF CALIFORNIA, BERKELEY. (2011)	DISCUSSES THE ELEMENTARY STEPS IN SUMMARIZATION OF DOCUMENT. IT ALSO DISCUSSES VARIOUS TECHNIQUES FOR EACH OF THE STEPS.

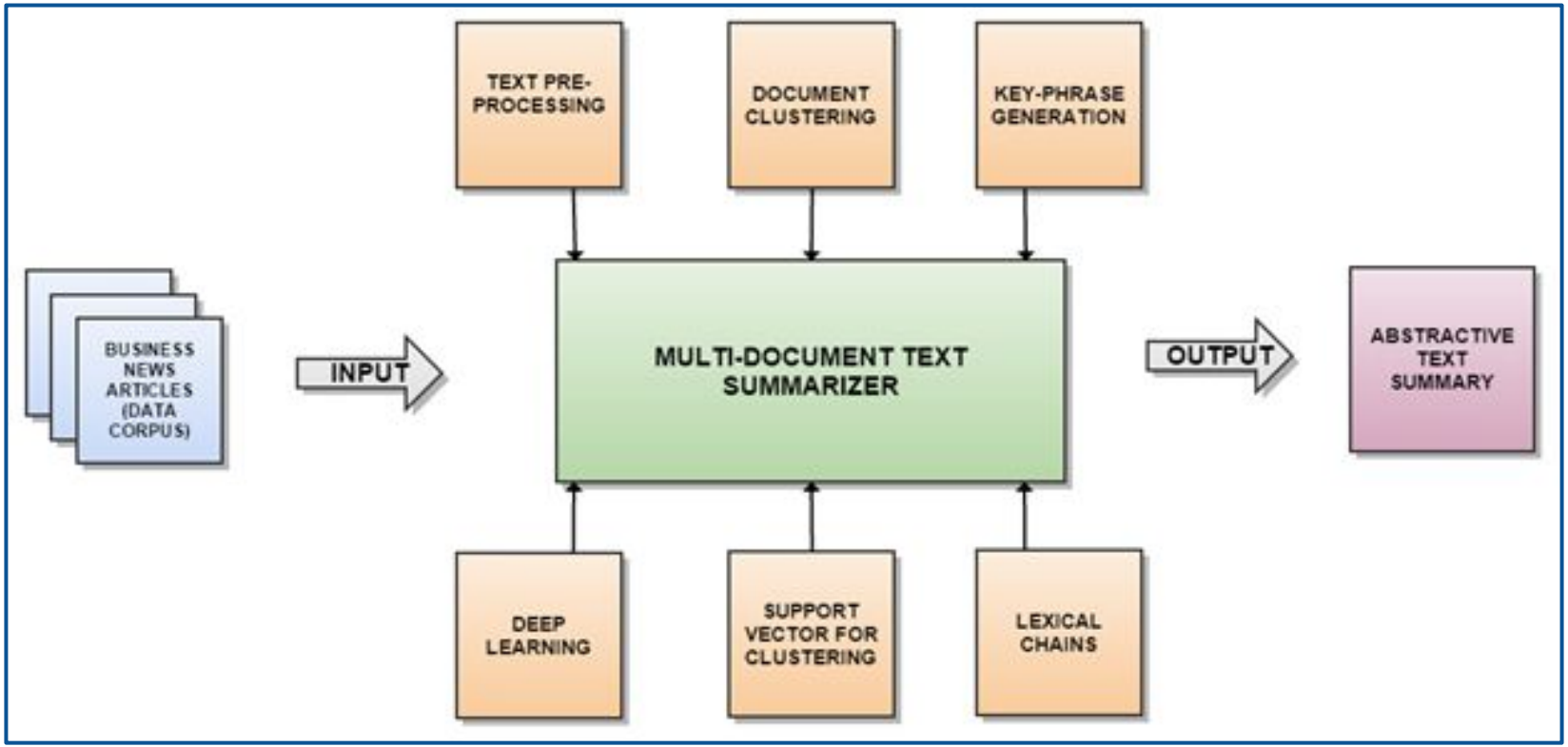
Literature Survey

TITLE	AUTHOR	PUBLICATIONS	DETAILS
AN APPROACH FOR TEXT SUMMARIZATION USING DEEP LEARNING ALGORITHM	PADMAPRIYA, G. AND K. DURAISWAMY	JOURNAL OF COMPUTER SCIENCE. (2014)	USES RESTRICTED BOLTZMANN MACHINE (RBM) ALGORITHM. THE SUMMARY IS GENERATED FOR DIFFERENT DOCUMENT SET FROM DIFFERENT KNOWLEDGE DOMAIN AND THE F-MEASURE IS CALCULATED
A SURVEY ON AUTOMATIC TEXT SUMMARIZATION	DIPANJAN DAS, ANDR`E F.T. MARTINS	SUBMITTED TO LANGUAGE TECHNOLOGIES INSTITUTE, CARNEGIE MELLON UNIVERSITY. (2007)	DISCUSSES SOME OF THE MOST RELEVANT APPROACHES BOTH IN THE AREAS OF SINGLE-DOCUMENT AND MULTIPLE-DOCUMENT SUMMARIZATION, GIVING SPECIAL EMPHASIS TO EMPIRICAL METHODS AND EXTRACTIVE TECHNIQUES

Literature Survey

TITLE	AUTHOR	PUBLICATIONS	DETAILS
SUPPORT VECTOR CLUSTERING	BEN-HUR, HORN, SIEGELMANN AND VAPNIK	JOURNAL OF MACHINE LEARNING RESEARCH. (2001)	DESCRIBES THE USAGE OF SUPPORT VECTORS (SUPERVISED) FOR CLUSTERING (UNSUPERVISED). RELEVANT MATHEMATICAL FORMULAE ARE OUTLINED.
USING LEXICAL CHAINS FOR TEXT SUMMARIZATION	REGINA BARZILAY AND MICHAEL ELHADAD	MATHEMATICS AND COMPUTER SCIENCE DEPT., BEN GURION UNIVERSITY OF NEGEV	DESCRIBES A LEXICAL CHAIN BASED TECHNIQUE FOR GENERATION OF SUMMARY

System at a Glance

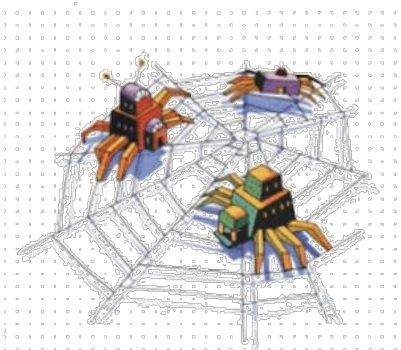


Steps of Summarization

- Data Corpus Building
- Document Clustering
- Data Preprocessing and Feature Matrix Building
- Generating Extractive Summary
- Generating Abstractive Summary
- Key-phrase Generation

Data Corpus Building

- Collects news articles with the help of rss feeds list and a crawling module.
- It returns a well formatted json for data corpus.
- Module used for crawling - FiveFilters.org
    - Input : Link to webpage containing the news article
    - Output : Full content stripped of ads and unwanted items.
- About 800 articles were collected over a period of 10 days.
- Module to build a complete set of data corpus (non-redundant data).



Data Corpus Building

- **Json Structure:-**

{ "root" :

[{ "title" : "Rise in sensex",

"content" : "The rise led..",

"pubdate" : "Mon, 26 Jan 2015..",

"newspaper" : "BBC",

"id" : 1},

{....},

...]

}

RSS Feeds :-

- Economic Times
- Reuters
- BBC
- The Economist
- Business Standards
- Forbes

Document Clustering

- K-means algorithm.
- Each document - represented as a tf-idf vector.
- Two levels of clustering :-
  - level 1 - 13 clusters
  - level 2 - 12 clusters
- Total 156 clusters
- Next ---->

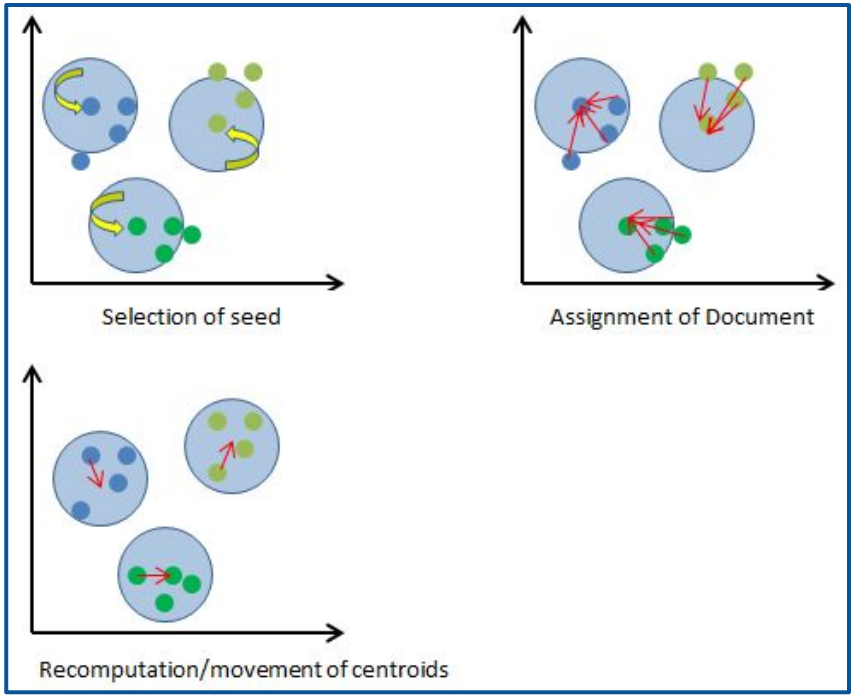


Fig : Document Clustering Using K-means

TF-IDF Matrix for Multi-Document

Term / Doc	Term1	Term2	Term3	Term4	Term5	….	TermN
Doc 1	0.67	0.92	0.11	0	0	...	0
Doc 2	0	0.13	0	0	0	...	0.23
Doc 3	0.45	0.31	0.09	0.31	0	...	0
...	...	...	...	...	...	...	...
Doc m	0	0	0	0.11	0.4	...	0

<---- Back

Data Preprocessing

- Sentence Tokenization
- Word Tokenization
- Parts Of Speech Tagging
- Named Entity Recognition
- Stemming
- Punctuation Removal

Feature Functions

- Sentence Length
- Term Frequency and Inverse Document Frequency (tf-idf)
- Title Similarity
- Location of Sentence
- Similarity with first sentence
- Domain specific features

Algorithms to Generate Summary

Deep Learning

- Performs preprocessing, constructs feature vector matrix which forms the input the RBM, and generates optimal feature vector set.
- Sentence with feature vector satisfying threshold values, are extracted.
- Restricted Boltzmann Machine

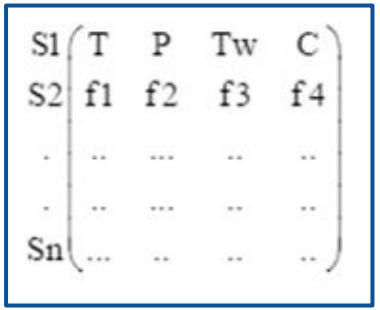


Fig : Feature Vector Matrix

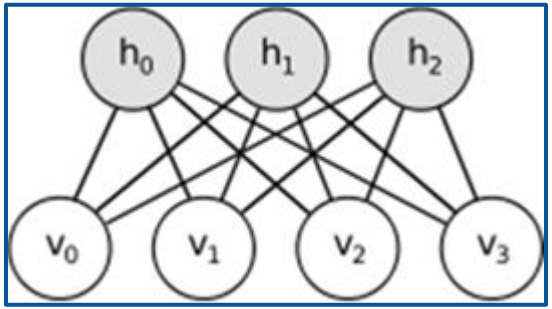
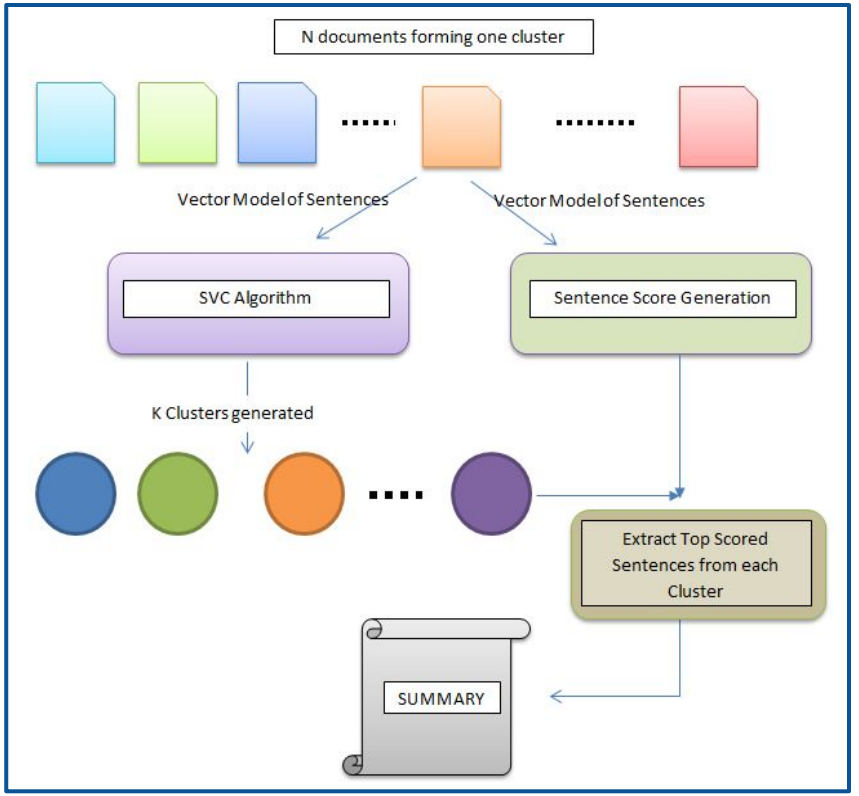


Fig : RBM

Support Vectors for Clustering

- Projection of data into higher dimension using rbf kernel function (2 parameters , C and gamma)
- Performs clustering of sentences, for each document cluster, using support vectors (SVC algorithm)
- Sentence scoring algorithm is then applied to obtain top scoring sentences from each cluster
- SVC Algorithm



Lexical Chains

- For summarization using Lexical Chains, the following steps are performed :
  - Sentences and words are tokenized and part of speech tagging is performed.
  - Lexical Chains are constructed, Lexical chains are scored and Significant sentences are then extracted.
- Generally, a procedure for constructing lexical chains follows three steps:
  - Select a set of candidate words.
  - For each candidate word, find an appropriate chain relying on a relatedness criterion among members of the chains
  - If it is found, insert the word in the chain and update it accordingly.
- Lexical Chains

Generating Abstractive Summary

- Pronoun Replacement

  eg. "John works for Google. He is enjoying in that company a lot."

  "John works for Google . John is enjoying in Google a lot ."

- Sentence Length
  - not too big or too small.
- Removal of Stress-words

  eg. [However, Furthermore, Hence, So, Also, Therefore, ...]

Key-phrase Generation

- Using python tool - RAKE
- Generates top four scoring key-phrases

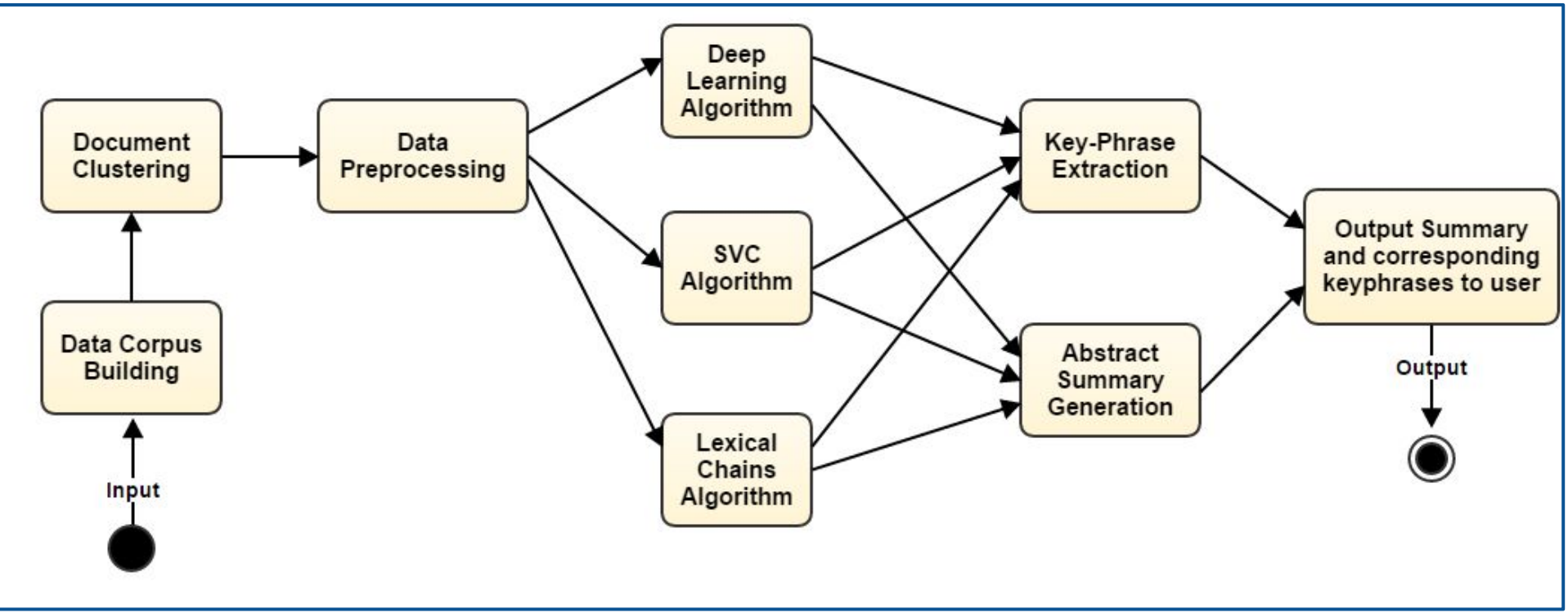
Tags : senior finance ministry official,wednesday asked bank chiefs,finance minister asked,percentage point reduction

Despite the signal, lenders appeared in no mood for an immediate reduction in rates with some bankers suggesting that the cut may not take place at the start of April.A senior finance ministry official, however, said it was enough of a signal.NEW DELHI: In what is being seen as a gentle prod to state-run lenders, the government on Wednesday asked bank chiefs to explain the reasons for their refusal to pass on gains from RBI's half a percentage point reduction in interest rate to borrowers.The finance minister asked us why rates were not being reduced, but there were no directive on cutting rates, a bank chief said after the meeting.

Evaluation of Summary

- Grammar
- Length
- Focus
- Coverage
- Well-Formatted
- Non-redundancy

Workflow of System



System Requirements

Hardware Requirements

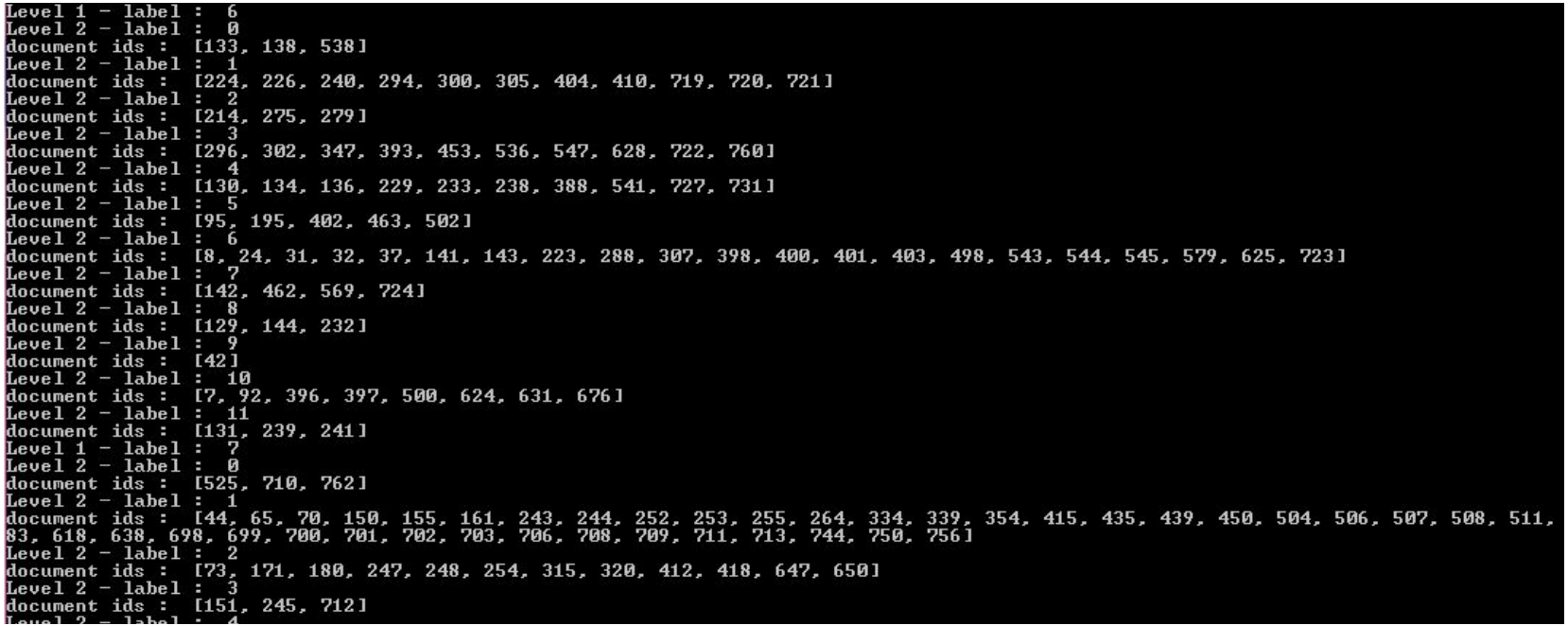
- Processor : Pentium Class PC (2.7 GHz or greater; faster processor or multiple processors recommended)
- Memory 4 GB of RAM or more recommended
- Operating system independent.

Software Requirements

- Python 2.6 or greater
- NumPy (>= 1.6)
- SciPy (>= 0.11)
- SciKit_Learn (>=0.15.2)
- NLTK (>= 3.0.1)
- Theano Library Latest Version
- Latest version of WAMP. We are using WAMP (2.5) with - Apache 2.4.9, PHP 5.5.12.

Snap-Shots

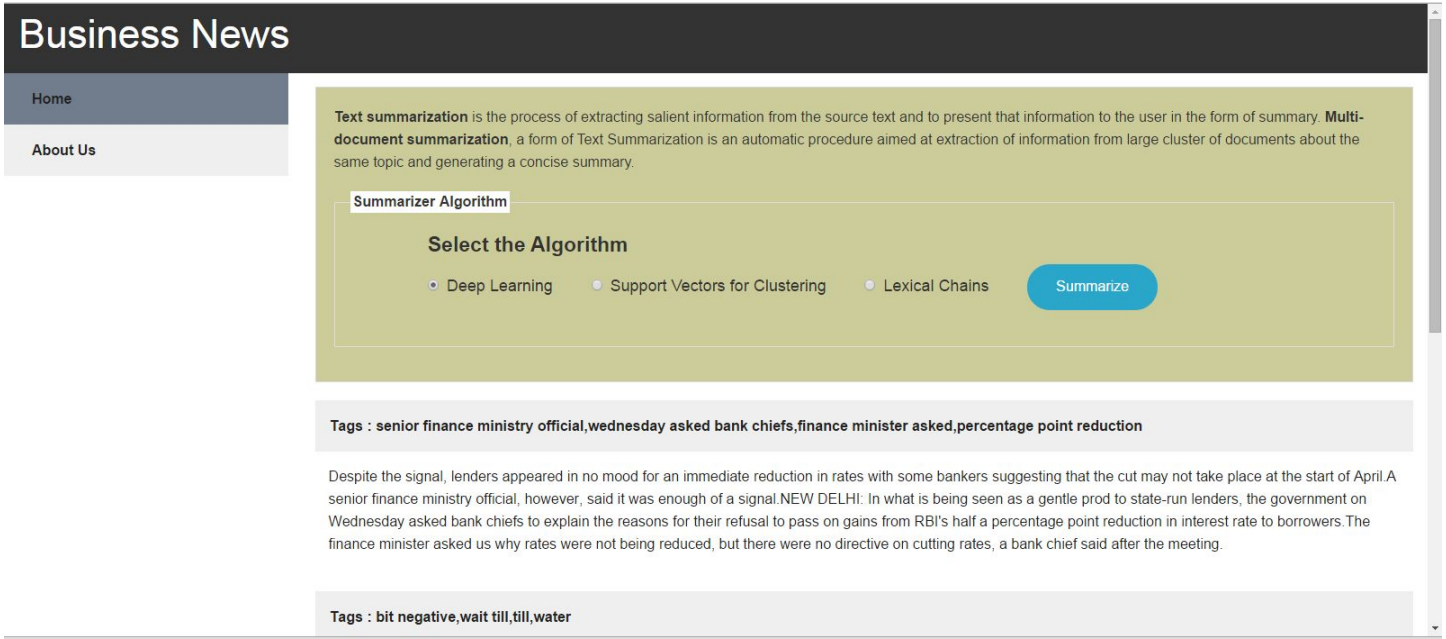
Screenshot - Document Clustering Output



```
Level 1 - label :  6
Level 2 - label :  0
document ids :  [133, 138, 538]
Level 2 - label :  1
document ids :  [224, 226, 240, 294, 300, 305, 404, 410, 719, 720, 721]
Level 2 - label :  2
document ids :  [214, 275, 279]
Level 2 - label :  3
document ids :  [296, 302, 347, 393, 453, 536, 547, 628, 722, 760]
Level 2 - label :  4
document ids :  [130, 134, 136, 229, 233, 238, 388, 541, 727, 731]
Level 2 - label :  5
document ids :  [95, 195, 402, 463, 502]
Level 2 - label :  6
document ids :  [8, 24, 31, 32, 37, 141, 143, 223, 288, 307, 398, 400, 401, 403, 498, 543, 544, 545, 579, 625, 723]
Level 2 - label :  7
document ids :  [142, 462, 569, 724]
Level 2 - label :  8
document ids :  [129, 144, 232]
Level 2 - label :  9
document ids :  [42]
Level 2 - label :  10
document ids :  [7, 92, 396, 397, 500, 624, 631, 676]
Level 2 - label :  11
document ids :  [131, 239, 241]
Level 1 - label :  7
Level 2 - label :  0
document ids :  [525, 710, 762]
Level 2 - label :  1
document ids :  [44, 65, 70, 150, 155, 161, 243, 244, 252, 253, 255, 264, 334, 339, 354, 415, 435, 439, 450, 504, 506, 507, 508, 511,
83, 618, 638, 698, 699, 700, 701, 702, 703, 706, 708, 709, 711, 713, 744, 750, 756]
Level 2 - label :  2
document ids :  [73, 171, 180, 247, 248, 254, 315, 320, 412, 418, 647, 650]
Level 2 - label :  3
document ids :  [151, 245, 712]
Level 2 - label :  4
```

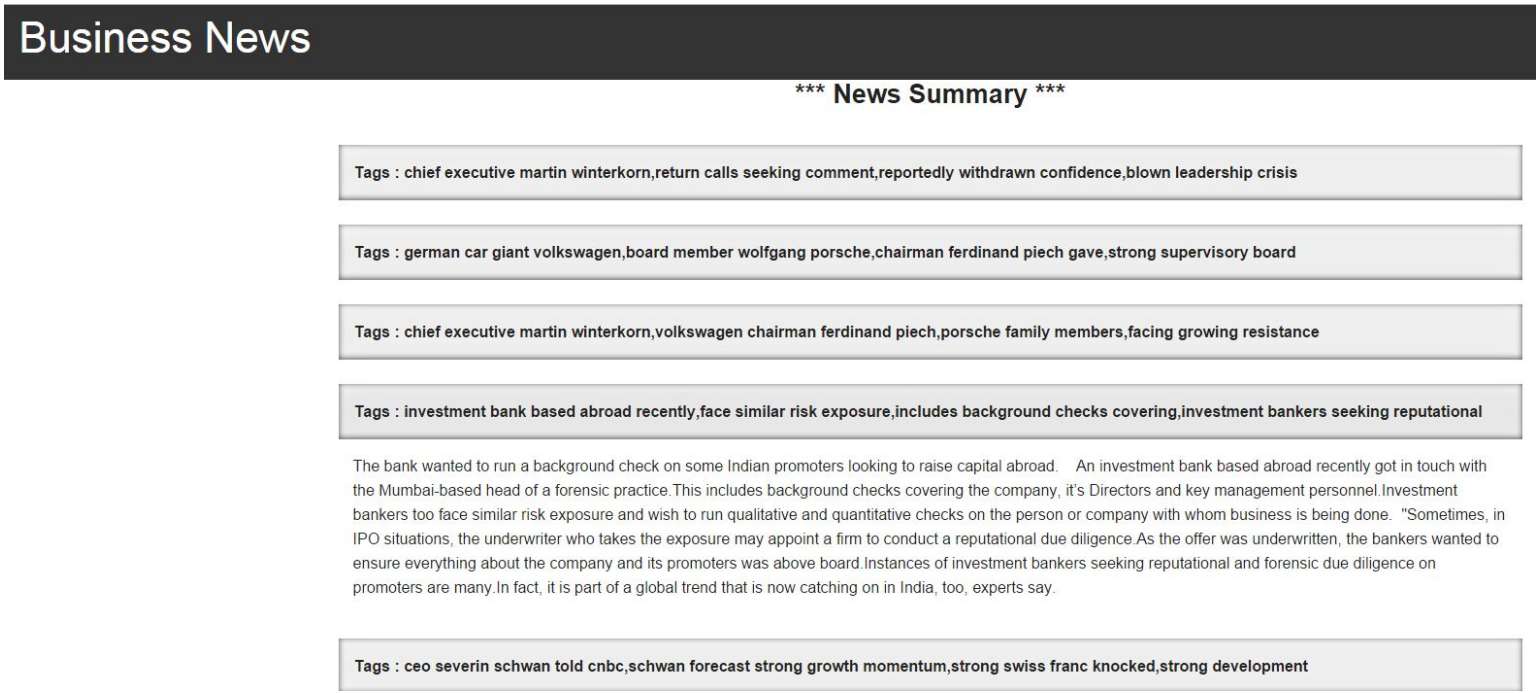
Snap-Shots

Screenshot - UI



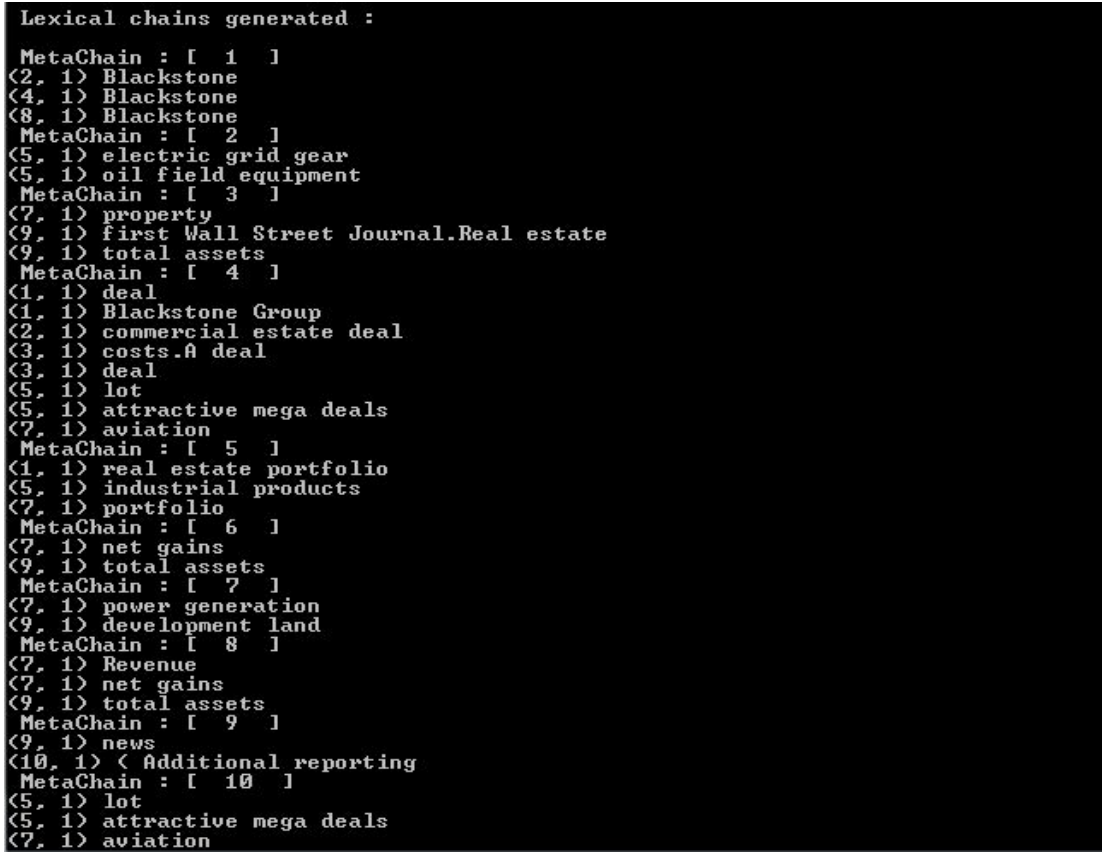
Snap-Shots

Screenshot - UI (News Tabs)



Snap-Shots

Screenshot - Lexical Chains



Snap-Shots

Screenshot - Deep Learning

```
 Training the RBM
Training epoch 0, cost is  -6.11595016693
Training epoch 1, cost is  -3.68571880059
Training epoch 2, cost is  -5.78013543574
Training epoch 3, cost is  -2.95275054728
Training epoch 4, cost is  -2.24370919272
Training epoch 5, cost is  -3.7845465245
Training epoch 6, cost is  -1.96619095172
Training epoch 7, cost is  -2.33138955942
Training epoch 8, cost is  -2.39460740346
Training epoch 9, cost is  -1.73847011621
Training epoch 10, cost is  -2.82747363719
Training epoch 11, cost is  -1.59509977476
Training epoch 12, cost is  -1.69160232963
Training epoch 13, cost is  -2.39254809383
Training epoch 14, cost is  -1.685824829
Training epoch 15, cost is  -2.61672907574
Training epoch 16, cost is  -1.31150167704
Training epoch 17, cost is  -1.08801981507
Training epoch 18, cost is  -2.49229570749
Training epoch 19, cost is  -1.68181083957
Training took 0.018443 minutes


 Sampling the RBM using Gibbs Sampling ...

 Optimal feature vector set :
[[ 0.24389781  0.90749953  0.86726153  0.01668702  0.94452642]
 [ 0.46217141  0.21888584  0.67861213  0.06526376  0.92083013]
 [ 0.24389781  0.90749953  0.86726153  0.01668702  0.94452642]
 [ 0.14361736  0.86300438  0.72296286  0.01025855  0.9528702 ]
 [ 0.17903178  0.81875683  0.92983255  0.21318344  0.63718684]]
```

Conclusion

- The project, “Abstractive Multi-Document Text Summarization“, undertaken was completed in the duration of about four months.
- The performance of this system was satisfactory and the results obtained were as desired and conforms to the specified requirements.
- We implemented “Deep Learning”, “Support Vectors for Clustering” and “Lexical Chain” to generate summaries.
- “Deep Learning” and “Support Vectors for Clustering” are our contributions to the field of Natural Language Processing specifically to Text Summarization.
- The system currently generates summary using a hybrid model (Extractive and Abstractive) of summarizer.

Future Work

- Feature Functions can be enhanced.
- Improvised algorithm for Document Clustering
- Acquire more servers and host algorithms to resolve memory constraints issue.

References

[1] Daniel Jacob Gillick, “The Elements of Automatic Summarization”, thesis submitted to University of California, Berkeley, 2011.

[2] Vasileios Hatzivassiloglou, Luis Gravano, Ankineedu MagantiAn, “Investigation of Linguistic Features and Clustering Algorithms for Topical Document Clustering”, Proc. of the 23rd ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'00), 2000.

[3] Kathleen R. McKeown, Regina Barzilay, David Evans, Vasileios Hatzivassiloglou, Judith L. Klavans, Ani Nenkova, Carl Sable, Barry Schiffman, Sergey Sigelman, “Tracking and Summarizing News on a Daily Basis with Columbia’s Newsblaster”, Department of Computer Science 450, Computer Science Building, Columbia University.

[4] Michael Steinbach George Karypis Vipin Kumar, “A Comparison of Document Clustering Techniques”, Department of Computer Science and Engineering, University of Minnesota.

References

[5] Vishal Gupta, “A Survey of Text Summarization Extractive Techniques”, Journal Of Emerging Technologies In Web Intelligence, Vol. 2, No. 3, August 2010.

[6] Dipanjan Das, Andre F.T. Martins, “A Survey on Automatic Text Summarization”, Language Institute of Technology, CMU November, 2007.

[7]Atif Khan, Naomie Salim, ”A Review On Abstractive Summarization Methods”, Journal of Theoretical and Applied Information Technology

[8] Jackie C U Cheung, “Comparing Abstractive and Extractive Summarization of Evaluative Text: Controversiality and Content Selection”, thesis submitted to University Of British Columbia, 2008

References

[9] PadmaPriya, G. and K. Duraiswamy, “An Approach For Text Summarization Using Deep Learning Algorithm”, Journal of Computer Science: 1-9, 2014.

[10] Ben-Hur, Horn, Siegelmann and Vapnik, “Support Vector Clustering”, Journal of Machine Learning Research 2 (2001): 125-137.

[11] Kees Jong, Elena Marchiori, Aad van der Vaart, “Finding Clusters using Support Vector Classifiers”, Department of Mathematics and Computer Science Free University Amsterdam The Netherlands.

[12] Regina Barzilay and Michael Elhadad, “Using Lexical Chains for Text Summarization”, Mathematics and Computer Science Dept., Ben Gurion University of Negev

References

[13] Alyona Medelyan, “https://www.airpair.com/nlp/keyword-extraction-tutorial”, website

[14] Ian H. Witten, Gordon W. Paynter, Eibe Frank, Carl Gutwin and Craig G. Nevill-Manning, “KEA: Practical Automatic Keyphrase Extraction”

[15] G ̈onen ̧c Ercan, “Automatic Text Summarization and Keyphrase Extraction”, a Master Degree thesis submitted to department of Computer Engineering, Bilkent University, September, 2006

Thank You !!

Appendix

Restricted Boltzmann Machine

- Contrastive Divergence is used for training the RBM.
- Markov chain initialised with data.
- k-steps of Gibbs sampling done.

Gibbs Sampling : -

$$h^{(n+1)} \sim sigm(W' v^{(n)} + c)$$
$$v^{(n+1)} \sim sigm(W h^{(n+1)} + b),$$

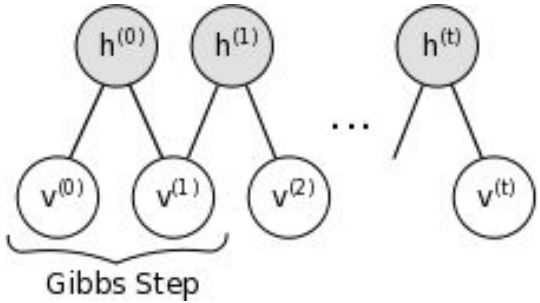


Fig : Markov Chain

←- Back

Support Vectors Clustering Algorithm

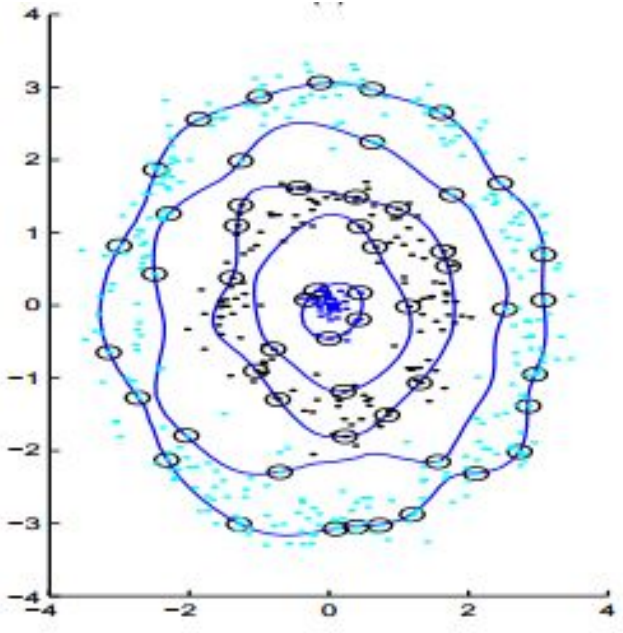


Fig : Nonlinear Data Separation

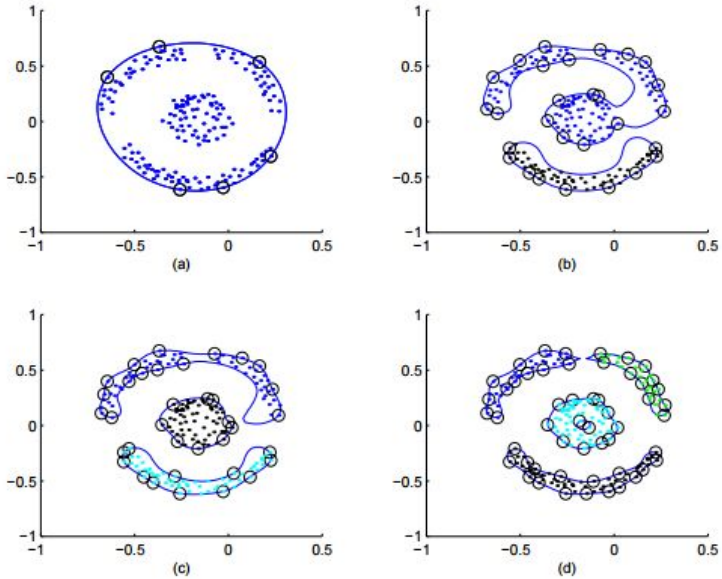


Fig : Formation of Clusters as gamma varies

←- Back

Lexical Chains

"***Mr***. *Kenny is the* ***person*** *that invented an* ***anesthetic machine*** *which uses* ***micro-computers*** *to control the rate at which an anesthetic is pumped into the blood…. "*

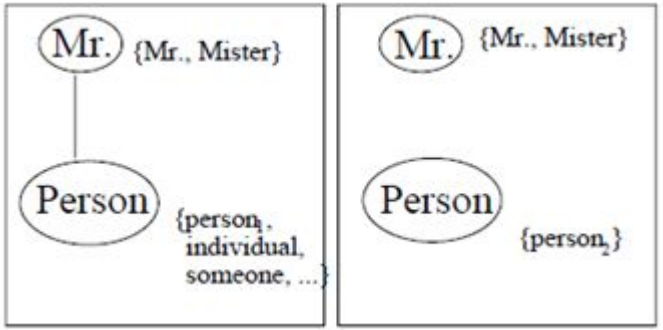


Fig : Step 1, Interpretation 1 and 2

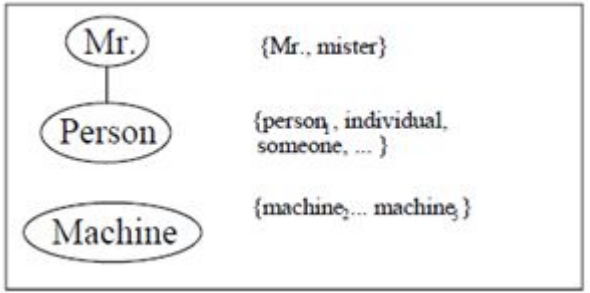


Fig : Step 2, Interpretation 1

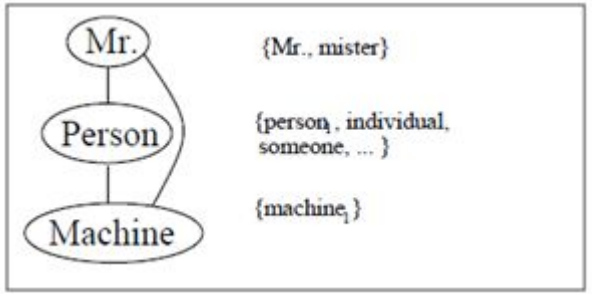


Fig : Step 2, Interpretation 2

←- Back