INTERNATIONAL JOURNAL OF STATISTICS AND MEDICAL INFORMATICS

Vol. 4 No. 1/2017, pp: 1-6

Natural Language Processing concepts and methods revisited

Editor, ISJMI

1. Introduction

Natural Language Processing [1,2] (NLP) evolved from the field computational linguistics which includes methods to study the language with the help of computers. It started with the process of translating languages [3] into another language with the use of machine translations algorithms. At next level, computers started understanding the language by parsing the sentences and deriving meaning out of the sentences. Moving further in this direction, rule based algorithms used in the next stage following statistical methods are used to process the language. Ambiguity in the natural languages is still a hurdle for the NLP systems even though lot work has been done to reduce the ambiguities. Statistical methods can help us to resolve ambiguities and learning from the set of data or corpus.

Usage of NLP systems in biomedical domain [4] is increasing as there is a need to understand the hidden knowledge in the vast text document present in the biomedical literature

2. Applications of NLP in Biomedical domain

NLP is used to extract information from the Electronic Medical Records [5], encoding of clinical documents [6], clinical decision support [7,8], and disease status identification [9,10] in combination with text mining

3. General Applications of NLP

NLP based applications are used in the fields of Machine Translation[3], dialog systems, Information retrieval, Information extraction, Named Entity Recognition, Question Answering and Sentiment Analysis[11].

4. Basic language terms and definitions used in Natural Language Processing

The Table-1 below provides commonly used basic terms and definitions in NLP systems which give the users fair idea about the system.

Vol. 4 No. 1/2017, pp: 1-6

Table 1 – Basic Terms and Definitions in NLP

Terms	Definitions
Semantics	Meaning of words
Morpheme	Sub part of the words which has meaning
Bag of words	Frequency of words
String of words	Linear sequence of words
Tree of words	Represented by recursive structure of language
Word Boundary	Space between words
Word formation	Inflection, Derivation and Compounding
Sentences boundary	Formed by full stop and semicolon
Syntax	Rules to form sentences from words(Grammar).
Parsing/Part of Speech tagging/Chunking	Dividing the sentences into parts using syntactic structure/Grammar
Text categorization	Assigning documents to predetermined list of topics
Lexicon	Dictionary or list of words
Tokenization	Divided the text into smaller units of words, numbers or punctuation
Discourse	Analysing the two or more connected sentences in a given text or document
Pragmatic	Analysing the text in context of world knowledge

5. Approaches to NLP [12]

i. Linguistic approach

Linguistic approach uses rule based approach wherein the set of rules are applied on given input and the rule which matches the condition is executed.

ii. Statistical approach

Statistical approach uses statistical models such as Hidden Markov Models wherein the models is defined by a set of states and its associated transition probabilities when the transition happens from state to another state. The states are hidden in the Models but output of each transition is observable with certain probabilities

6. Text Classification[13]

6.1 Supervised Classification

During the training phase a set of features is extracted from the input and then label is assigned to each feature which acts as a classifier. Once the new input text is fed into the classifier and a set of features are extracted from the given new input. Now classifier assigns

Vol. 4 No. 1/2017, pp: 1-6

label to new input using the training set and the same is added into training set for future classifications.

Examples of supervised classification can be seen in assigning topics to given input or labelling a given text input as positive or negative sentiments or classifying an email as spam or not spam

6.1.1. Naïve Bayes classifier

Naïve Bayes classifier [13] uses prior probabilities to assign label to a given input through the features and associated feature weight for the each label in the training set.

For example consider a case when an article which needs to classified into a particular topic using the Naïve Bayes classifier. Our training classifier contains documents related to diseases such as cancer, diabetes mellitus, and arterial disease, feature set and corresponding labels. If the input article is on cancer then the classifier starts extracting features from the given article which contains more features related to cancer and more weight will be given to the features related to the cancer during the classification process and corresponding prior probabilities calculated based on the feature set which will make classifier to assign the label cancer to the given article

6.1.2. Decision Tree

Decision tree[14] contains nodes and leaf wherein nodes represent the classifying conditions and leaf represent classifying labels. The labels are selected based on the classifying conditions satisfied at the node stage. The process starts with the input of text input from the root node and the given text input is split into two at the first stage. This process is repeated till the leaf cannot be further split into. The classifying conditions may be made up of presence of absence of words, similarity or dissimilarity of words in the document. The Decision tree also uses Part of Speech tagging to create nodes and leafs.

6.1.3. Support Vector Machine

Support Vector Machine [15] techniques consider the given text input as multidimensional space wherein the input is scattered on the multidimensional space. SVM divide the space into hyperplanes (n-1 sub planes in n dimensional plane) and those hyperplanes are selected which minimize the distances from the words within the planes and but maximises the distance between the planes.

For example let us consider a two dimensional space such which contains positives and negatives words. SVM divides the space into hyperplanes (here it is as two disconnected subspace which are separated by a line) and calculates the distance from the hyperplane for each word. Then the task is to select the hyperplane which minimizes the distance for each

Vol. 4 No. 1/2017, pp: 1-6

word from the subspace/line and maximizes the distances between subspaces. Words which are closer to the hyperplane/line form the support vector for that hyperplane or subspace.

6.2 Unsupervised methods[1]

6.2.1 Clustering methods

To group the given input text into clusters so that words in the same cluster are similar and different clusters are dissimilar. The clustering can be done in two ways hierarchical (agglomerative) clustering and partitioning clustering (k-means clustering). In the clustering method Term Frequency-Inverse Document Frequency (IF-DFT) algorithm (weight of a term is proportional to the number of times the term appears in each document) is used to create the clusters

7. Language Modelling [1,16]

Language modelling involves assigning probability to words and sentences. It helps to predict the sequence of words which are likely to be present in a text input. Language models deals with sparse or unknown words through the use of Smoothing algorithm

7.1. Zipf's Law[17]

Zipf's Law is an important concept in the language modelling which deals with the words frequency and its weight or rank in a corpus.

When word are ranked by its frequency in a given text input, then frequency * rank of the frequency is equal to a constant

i.e if frequency of word is f and rank of the words frequency is r then

$f*r = c$

7.2. Hidden Markov Models[18]

Hidden Markov Models is defined by a set of states and its associated transition probabilities and output or emission probabilities. Transition probabilities are calculated when the transition happens from state to another state. The states are hidden in the Models but output of each transition is observable with certain probabilities

7.3. n-grams models[19]

An ngram is a continuous sequence of n items in a text where one can predict the nth word from the previous words. 3gram represents continuous sequence of 3 words wherein we can predict the third word from the previous 2 words

Vol. 4 No. 1/2017, pp: 1-6

8. Information Retrieval and NLP

The starting phase of an automated NLP system involves information retrieval. Some of the information retrieval concepts related to NLP are discussed below

8.1. Latent Semantic Analysis (LSA) and Latent Semantic Indexing (LSI) [20]

In information retrieval process text input is treated as a Document Term matrix wherein rows represents terms, column represents documents. TF_IDF is used to represents the weight of the terms in the text input.

Latent Semantic Analysis (LSA) is used to uncover the meaning of the text in response to the queries during the information retrieval process. LSA uses Singular Value Decomposition techniques to reduce the dimensionality of the Document-Term matrix by grouping the similar words in given text.

9. Software to carry out NLP tasks

The following table gives the list of software available to carry out NLP tasks

Table-2 List of Software

Name of the software	Type	URL
Standford NLP[21]	Open source	www.nlp.stanford.edu
NLTK with python[22]	Open source	www.python.org www.nltk.org
R Statistical Package and R-Studio 1. OpenNLP 2. tm 3. rNLP	Open source	https://www.r-project.org/ https://www.rstudio.com/products/rpackages/

10. Conclusion

The paper revisited the concepts and methods of Natural Language Processing Systems.

11. References

1. Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing* (Vol. 999). Cambridge: MIT press.
2. Liddy, E. D. (2001). Natural language processing.
3. Hutchins, W. J. (1986). *Machine translation: past, present, future* (p. 66). Chichester: Ellis Horwood.
4. Spyns, P. (1996). Natural language processing. *Methods of information in medicine*, *35*(4), 285-301.

Vol. 4 No. 1/2017, pp: 1-6

5. Cronin, T. (2014). Automation of Medical Record Risk Factor Tagging Using Machine Learning and Natural Language Processing Methods.
6. Friedman, C., Shagina, L., Lussier, Y., & Hripcsak, G. (2004). Automated encoding of clinical documents based on natural language processing. *Journal of the American Medical Informatics Association*, *11*(5), 392-402.
7. Demner-Fushman, D., Chapman, W. W., & McDonald, C. J. (2009). What can natural language processing do for clinical decision support?. Journal of biomedical informatics, 42(5), 760-772.
8. Szlosek, D. A., & Ferrett, J. (2016). Using Machine Learning and Natural Language Processing Algorithms to Automate the Evaluation of Clinical Decision Support in Electronic Medical Record Systems. *eGEMs*, *4*(3).
9. Alemzadeh, H., & Devarakonda, M. (2017, February). An NLP-based cognitive system for disease status identification in electronic health records. In *Biomedical & Health Informatics (BHI), 2017 IEEE EMBS International Conference on* (pp. 89-92). IEEE.
10. Zeng, Q. T., Goryachev, S., Weiss, S., Sordo, M., Murphy, S. N., & Lazarus, R. (2006). Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system. *BMC medical informatics and decision making*, *6*(1), 30.
11. Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, *5*(4), 1093-1113
12. Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P. (2011). Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, *12*(Aug), 2493-2537.
13. Go, A., Bhayani, R., & Huang, L. (2009). Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, *1*(2009), 12.
14. Brill, E. (1995). Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging. *Computational linguistics*, *21*(4), 543-565.
15. Collobert, R., & Weston, J. (2008, July). A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning* (pp. 160-167). ACM.
16. Winograd, T. (1972). Understanding natural language. *Cognitive psychology*, *3*(1), 1-191.
17. Piantadosi, S. T. (2014). Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic bulletin & review*, *21*(5), 1112-1130.
18. Huang, X. D., Ariki, Y., & Jack, M. A. (1990). *Hidden Markov models for speech recognition* (Vol. 2004). Edinburgh: Edinburgh university press.
19. Brown, P. F., Desouza, P. V., Mercer, R. L., Pietra, V. J. D., & Lai, J. C. (1992). Class-based n-gram models of natural language. *Computational linguistics*, *18*(4), 467-479.
20. Landauer, T. K. (2006). *Latent semantic analysis*. John Wiley & Sons, Ltd.
21. Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J. R., Bethard, S., & McClosky, D. (2014, June). The stanford corenlp natural language processing toolkit. In *ACL (System Demonstrations)* (pp. 55-60).
22. Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".