IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727,
PP 51-55
www.iosrjournals.org

Similarities in words Using Different Pos Taggers

Kalpana B. Khandale[1],Ajitkumar Pundage[2],C. Namrata Mahender[3]

[1]*(Department Of CS & IT, Dr. B.A.M.U. Aurangabad, India College/ University Name, Country Name)*
[2]*(Department Of CS & IT, Dr. B.A.M.U. Aurangabad, India College/ University Name, Country Name)*
[3]*(Department Of CS & IT, Dr. B.A.M.U. Aurangabad, India College/ University Name, Country Name)*

Abstract: *In the research area of the computational linguistic, there are the vast varieties of text data available and there is need to sort out it. Part-of-speech tagging is one of the most important part of the natural language processing which help us to identify the proper tag for the given text or sentences. This paper presents the basic techniques using four different part-of-speech tagger (POS tagger). With these tools we have found the differences among the tagged word in different way. From the four tools we have seen the different result for the same word.*

Keywords: *Question Answering System, NLTK, Freeling, Cognitive POS tagger, NLP tagger*

I. Introduction

1.1 Question Answering System:

In the vast variety of text data available in the question answering system and there are various problems faced by the researcher and “question understanding” is one of the problem of the QA system. With the help of part-of-speech tagger we found the proper tag of the word and understand it properly either it is noun, verb, and adjective and so on. This paper presents the different part of speech tagging of the question sentences. This paper presents the four different tools for the part-of speech tagging. And try to understand the tagging of the same word in four different ways.

Question answering (QA) is a computer science discipline within the fields of information retrieval and natural language processing (NLP), which is concerned with building systems that automatically answer questions posed by humans in a natural language. There are three component of QA system that are Question processing, Document and passage selection and Answer extraction.

1.2 application of qa:

There is several application of the question answering system. We have use question answering in various field. Information retrieval [1], Online Examination System [2], Online Business strategy [3], Language processing by computers [4], Machine translation [5], Human and machine interaction [6], Speech synthesis[7], Natural language processing [6], Language learning [6], Intelligent word processing [8], Document management [9], Text categorization/ summarization [10] etc.

2. Pos Tagging:

The process of assigning one of the parts of speech to the given word is called Parts Of Speech tagging. It is commonly referred to as POS tagging. Parts of speech include nouns, verbs, adverbs, adjectives, pronouns, conjunction and their sub-categories. Parts Of Speech tagger or POS tagger is a program that does this job. Taggers use several kinds of information: dictionaries, lexicons, rules, and so on. Dictionaries have category or categories of a particular word. That is a word may belong to more than one category. For example, run is both noun and verb. Taggers use probabilistic information to solve this ambiguity. [11]

Tagset is the set of tags from which the tagger is supposed to choose to attach to the relevant word. Every tagger will be given a standard tagset. The tagset may be coarse such as NN (Noun), VB (Verb), JJ (Adjective), RB (Adverb), IN (Preposition), and CC (Conjunction) and so on.

2.1 Architecture Of Pos Tagger

2.1.1 Tokenization: The given text is divided into tokens so that they can be used for further analysis. The tokens May Be Words, Punctuation Marks, And Utterance Boundaries.

2.1.2 Ambiguity Look-Up: This is to use lexicon and a guessor for unknown words. While lexicon provides list of word forms and their likely parts of speech, guessors analyze unknown tokens.

2.1.3 Ambiguity Resolution: This is also called disambiguation. Disambiguation is based on information about word such as the probability of the word. For example, power is more likely used as noun than as verb.

3. Tools Use For The Pos Tagging:

In this paper we have use four tools for the pos tagging. Those are NLTK, FREELING, NLP and POS tagger. In these four tools we found some differences on the basis of the ranking of the question sentence. Here also we have found the differences within the question sentence. In table 1 shows the difference between four tools and same word tagged as different way.

3.1 Natural Language Toolkit

The Natural Language Toolkit, or more commonly NLTK, is a suite of libraries and programs for symbolic and statistical natural language processing (NLP) for English written in the Python programming language. It is created by the Department of Computer Science and Information Science at the University of Pennsylvania.

NLTK is intended to support research and teaching in NLP or closely related areas, including empirical linguistics, cognitive science, artificial intelligence, information retrieval, and machine learning. NLTK supports classification, tokenization, and stemming, tagging, parsing, and semantic reasoning functionalities. The Natural Language Toolkit (NLTK) is a platform used for building Python programs that work with human language data for applying in statistical natural language processing (NLP). It contains text processing libraries for tokenization, parsing, classification, stemming, tagging and semantic reasoning. It also includes graphical demonstrations and sample data sets as well as accompanied by a cook book and a book which explains the principles behind the underlying language processing tasks that NLTK supports. [12]

3.2 Freeling:

FreeLing is a C++ library providing language analysis functionalities (morphological analysis, named entity detection, PoS-tagging, parsing, Word Sense Disambiguation, Semantic Role Labeling, etc.) for a variety of languages (English, Spanish, Portuguese, Italian, French, German, Russian, Catalan, Galician, Croatian, Slovene, among others). FreeLing is designed to be used as an external library from any application requiring this kind of services. Nevertheless, a simple main program is also provided as a basic interface to the library, which enables the user to analyze text files from the command line. [13]

3.3 Cognitive Pos Tagger:

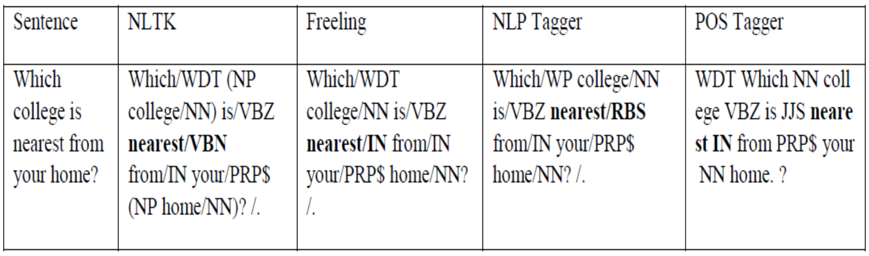
This requires programs that can, at some level, understand natural language text – categorizing documents by topic or function, retrieving entities and concepts – rather than looking for specific strings in text. Some of the demos below directly address such Information Extraction tasks (e.g. Named Entity Recognition, Data less Classification), while others exhibit fundamental natural language technologies that can support higher-level applications (e.g. Part-of-Speech Tagging, Shallow Parsing, Semantic Role Labeling). [14]

3.4 Nlp Tagger:

The NLP tagger provides following services to linguistic users:

The tokenization service splits any piece of text into a list of tokens. Based on your option, a token can be a word, a punctuation mark, a symbol, or a number. The part-of-speech tagging service goes one step further by providing the most appropriate part-of-speech to each of the token. Be aware that part-of-speech in this sense is much broader than the categories we learned in traditional grammar books. For the complete part-of-speech, i.e. the tags used by this tagger. The morphological analysis service analyzes the morphological processes a word has undergone to get its current form from its base form. [15]

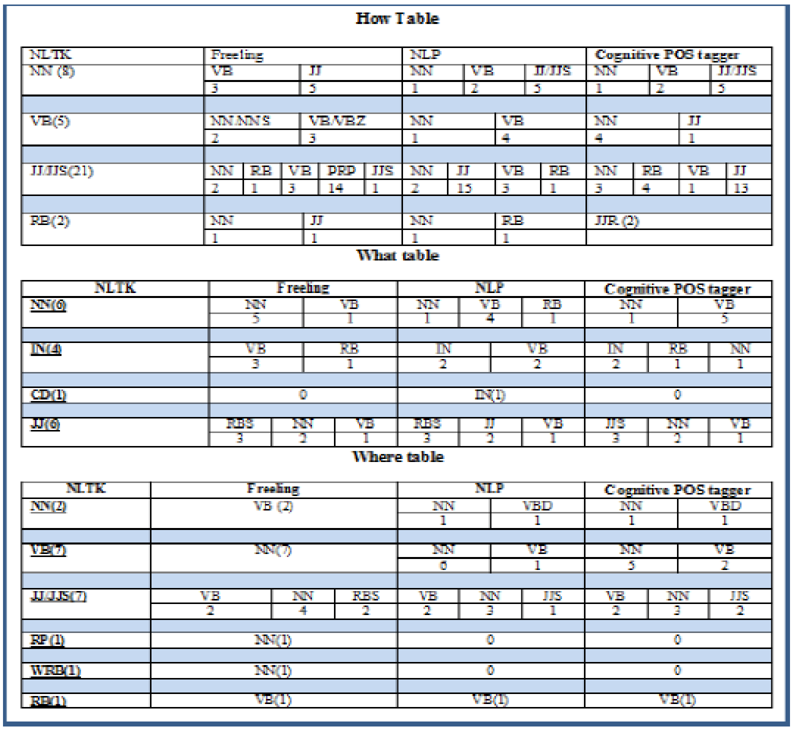
Table 1. Difference between four tools



Sentence	NLTK	Freeling	NLP Tagger	POS Tagger
Which college is nearest from your home?	Which/WDT (NP college/NN) is/VBZ **nearest/VBN** from/IN your/PRP$ (NP home/NN)? /.	Which/WDT college/NN is/VBZ **nearest/IN** from/IN your/PRP$ home/NN? /.	Which/WP college/NN is/VBZ **nearest/RBS** from/IN your/PRP$ home/NN? /.	WDT Which NN coll ege VBZ is JJS **neare st IN** from PRP$ your NN home. ?

In the sentence, “Which college is nearest from your home?”

In which we can see the difference between the tagged word that are in NLTK "nearest" tagged as **BN**, in Freeling same word tagged as **IN**, NLP tagger same tagged as the **RBS** and in POS tagger it tagged as **JJS** The bold words show the difference of the tagged word from these four tools.



How Table

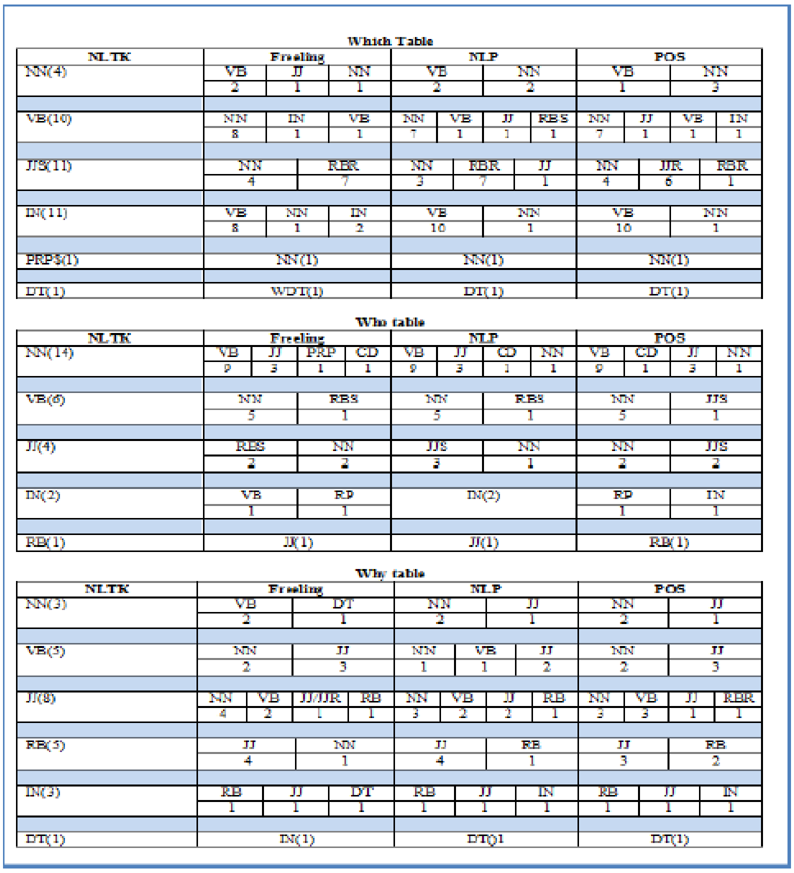
NLTK	Freeling	NLP	Cognitive POS tagger
NN (8)	VB 3 JJ 5	NN 1 VB 2 JJ/JJS 5	NN 1 VB 2 JJ/JJS 5
VB(5)	NN/NNS 2 VB/VBZ 3	NN 1 VB 4	NN 4 JJ 1
JJ/JJS(21)	NN 2 RB 1 VB 3 PRP 14 JJS 1	NN 2 JJ 15 VB 3 RB 1	NN 3 RB 4 VB 1 JJ 13
RB(2)	NN 1 JJ 1	NN 1 RB 1	JJR (2)

What table

NLTK	Freeling	NLP	Cognitive POS tagger
NN(6)	NN 5 VB 1	NN 1 VB 4 RB 1	NN 1 VB 5
IN(4)	VB 3 RB 1	IN 2 VB 2	IN 2 RB 1 NN 1
CD(1)	0	IN(1)	0
JJ(6)	RBS 3 NN 2 VB 1	RBS 3 JJ 2 VB 1	JJS 3 NN 2 VB 1

Where table

NLTK	Freeling	NLP	Cognitive POS tagger
NN(2)	VB (2)	NN 1 VBD 1	NN 1 VBD 1
VB(7)	NN(7)	NN 6 VB 1	NN 5 VB 2
JJ/JJS(7)	VB 2 NN 4 RBS 2	VB 2 NN 3 JJS 1	VB 2 NN 3 JJS 2
RP (1)	NN(1)	0	0
WRB(1)	NN(1)	0	0
RB(1)	VB(1)	VB(1)	VB(1)



Which Table

NLTK	Freeling	NLP	POS
NN(4)	VB 2 JJ 1 NN 1	VB 2 NN 2	VB 1 NN 3
VB(10)	NN 8 IN 1 VB 1	NN 7 VB 1 JJ 1 RBS 1	NN 7 JJ 1 VB 1 IN 1
JJS(11)	NN 4 RBR 7	NN 3 RBR 7 JJ 1	NN 4 JJR 6 RBR 1
IN(11)	VB 8 NN 1 IN 2	VB 10 NN 1	VB 10 NN 1
PRP$(1)	NN(1)	NN(1)	NN(1)
DT(1)	WDT(1)	DT(1)	DT(1)

Who table

NLTK	Freeling	NLP	POS
NN(14)	VB 9 JJ 3 PRP 1 CD 1	VB 9 JJ 3 CD 1 NN 1	VB 9 CD 1 JJ 3 NN 1
VB(6)	NN 5 RBS 1	NN 5 RBS 1	NN 5 JJS 1
JJ(4)	RBS 2 NN 2	JJS 3 NN 1	NN 2 JJS 2
IN(2)	VB 1 RP 1	IN(2)	RP 1 IN 1
RB(1)	JJ(1)	JJ(1)	RB(1)

Why table

NLTK	Freeling	NLP	POS
NN(3)	VB 2 DT 1	NN 2 JJ 1	NN 2 JJ 1
VB(5)	NN 2 JJ 3	NN 1 VB 1 JJ 2	NN 2 JJ 3
JJ(8)	NN 4 VB 2 JJ/JJR 1 RB 1	NN 3 VB 2 JJ 2 RB 1	NN 3 VB 3 JJ 1 RBR 1
RB(5)	JJ 4 NN 1	JJ 4 RB 1	JJ 3 RB 2
IN(3)	RB 1 JJ 1 DT 1	RB 1 JJ 1 IN 1	RB 1 JJ 1 IN 1
DT(1)	IN(1)	DT()1	DT(1)

Fig 1. Difference of the pos tagging for the same words in four different tools

The above figure shows the four different tools for the POS tagging. With the help of this we have observe the differences of the same words have different tag. As compare to NLTK tool the other three tools are tagged the same word different way. Bold and underlined word shows tagged words from the NLTK tool. But from the whole four POS tagger there is the NLTK gives the better result. In the table the number of part-of-speech tag counted and it compared with other three tools and observes the different tagging of the same word. For example, from the 28 sentences of the "HOW" type of questions tagged in NLTK tool and the noun (NN) is occurred eight times in those sentences. This NOUN tag is different into the other three tools such as in Freeling on the place of NOUN, the (VB) that is verb occurred as 3 times and adjective (JJ) occurred as 5 times. Likewise in the NLP tagger on the place of NOUN, VB occurred as 2 times, JJ as 5 times and only one time the NN tag is placed. In the Cognitive POS tagger on the place of NOUN the tagged word same like the NLP

IV. Methodology

4.1 Tokenization:

In this type of method we first select sentence from the data set and tokenize it into the words. In simple way "tokenization is the techniques which is help to decompose the sentences into the word."

4.2 Pos Tag:

The process of assigning one of the parts of speech to the given word is called Parts Of Speech tagging. It is commonly referred to as POS tagging. Parts of speech include nouns, verbs, adverbs, adjectives, pronouns, conjunction and their sub-categories. With the four types of tools have use to identify the correct tagging of the word. Those are the NLTK, Freeling, NLP and the Cognitive POS tagger.

V. Conclusion

Question answering (QA) is a computer science discipline within the fields of information retrieval and natural language processing (NLP). There are three component of the question answering system that are, question processing, document or passage selection and answer extraction. QA system has many applications the like HCI, NLP, text categorization and summarization etc. For the every field of QA the POS tagger is most important to know what the tag of the word is. In our work we have work on the four different tools for the POS tagging, NLTK, FREELING, NLP tagger and Cognitive POS tagger. There we found few differences in same words. From the 350 wh-type question sentences, total 154 sentences are tagged in different way with the four different tools. Under which the part of speech like NN, VB, JJ, and RB etc. are tagged different way for the same word.

References

[1]. Rodrigo, Á, Perez-Iglesias, J., Peñas, A., Garrido, G., & Araujo, L. (2010). A Question Answering System based on Information Retrieval and Validation. In CLEF (Notebook Papers/LABs/Workshops).

[2]. Dhokrat, A., Gite, H. R., & Mahender, C. N. (2012). Computation Linguistic: Online Subjective Examination Modeling. Advances in Computational Research, ISSN, 0975-3273.

[3]. Fleischman, M., Hovy, E., & Echihabi, A. (2003, July). Offline strategies for online question answering: Answering questions before they are asked. In Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1 (pp. 1-7).

[4]. Association for Computational Linguistics. Gupta, P., & Gupta, V. (2012).A survey of text question answering techniques. International Journal of Computer Applications, 53(4).

[5]. Espana-Bonet, C., & Comas, P. R. (2012, April).Full machine translation for factoid question answering. In Proceedings of the Joint

[6]. Workshop on Exploiting Synergies between Information Retrieval and Machine Translation (ESIRMT) and Hybrid Approaches to

[7]. Machine Translation (HyTra) (pp. 20-29). Association for Computational Linguistics.

[8]. Bilotti, M. W., & Nyberg, E. (2006).Evaluation for scenario question answering systems. In Proceedings of the International Conference on Language Resources and Evaluation.

[9]. Trilla, A. (2009). Natural Language Processing techniques in Text-To-Speech synthesis and Automatic Speech Recognition.Departament

[10]. de Tecnologies Media Enginyeria i Arquitectura La Salle (Universitat Ramon Llull), Barcelona, Spain atrilla@ salle.url. edu.

[11]. Aggarwal, M. (2011). Information Retrieval and Question Answering NLP Approach: An Artificial Intelligence Application. International Journal of Soft Computing and Engineering (IJSCE) ISSN, 1(11), 43-45.

[12]. Amato, F., Mazzeo, A., Penta, A., &Picariello, A. (2008, September).Using NLP and ontologies for notary document management systems.In2008 19th International Workshop on Database and Expert Systems Applications (pp. 67-71).IEEE.

[13]. Budzik, J., & Hammond, K. J. (1998). Learning for Question Answering and Text Classification: Integrating Knowledge-Based and Statistical Techniques. In AAAI Workshop on Text Classification.

[14]. Wang, M. (2006). A survey of answer extraction techniques in factoid question answering. Computational Linguistics, 1(1).

[15]. Brill, E. (1992, February). A simple rule-based part of speech tagger. In Proceedings of the workshop on Speech and Natural Language (pp. 112-116).

[16]. Bird, S. (2006, July). NLTK: the natural language toolkit. In Proceedings of the COLING/ACL on Interactive presentation sessions (pp. 69-72). Association for Computational Linguistics.

[17]. Carreras, X., Chao, I., Padró, L., & Padró, M. (2004, May). FreeLing: An Open-Source Suite of Language Analyzers. In LREC. http://cogcomp.cs.illinois.edu/page/demo_view/pos http://nlpdotnet.com/services/Tagger.aspx

[18]. Ajitkumar M. Pundge, Khillare S.A, Dr. C. Namrata Mahender, "Question Answering System, Approaches and Techniques: A Review", International Journal of Computer Applications (0975 – 8887)Volume 141 – No.3, May 2016