Comparison of Sentences using POS Tagging Tool under Subjective Examination

Aarti P Raut[1*†], **C Namrata Mahender**[1‡]

[1] *Department of computer science and IT, Dr.B.A.M University, Aurangabad, India.*

Abstract

The Question answering system is used to generate the correct result that is asked by humans in natural language. In an online examination system, most of the work has been done but still, problems occur in preprocessing i.e. Part Of Speech (POS). POS tagger is used to properly tag each word in the sentences. In this paper, we used two data sets i.e. TREC DATA and data collected from the student. We apply the POS tagger to both data sets and compare the result. For generating the POS tagger we used NLTK and spaCy libraries for comparison. We observed that using those libraries the same word has a different tag. Using both tools, we computed the difference between the words and assigned the count to the POS tagging on that result, we calculate the accuracy of both libraries. The result shows that the spaCy library is best for POS tagging because it generates more correct results as compared to NLTK.

Keywords: Question Answering System (QAS), Part Of Speech (POS) Tagger, TREC, NLTK, Spacy

1 Introduction

Question Answering Systems (QAS) are automated methods for retrieving precise responses to questions in natural language. This system is a reliable technology that can use a structured database or a collection of data that are written in natural language to automatically respond to queries posed by humans in that language. A large amount of text data is available on the question-answering system. The understanding sentence is one of the important issues in the QA system. Part of speech tagger is one of the preprocessing techniques that are applied

[*] Corresponding author.
[†] E-mail: rautaarti443@gmail.coma
[‡] E-mail: cnamrata.csit@bamu.ac.in

to the sentence to understand the type of each word. For understanding the sentence first to know the type of each word so we used the part of speech tagger (POS tagger) which is crucial to understand the words' meaning. A grammatical category known as a part of speech includes types like verbs, nouns, adjectives, adverbs, determiners, etc., and their subtypes. For the Grammatical word used the tag set for each type like Noun (NN), Proper Noun (NNP), etc.

Information retrieval (1), online examination systems (2), machine translation (3), text summarization (4), etc. are the important application of QA systems. During the pandemic, the online examination system was used to conduct online exams. Even though the online examination system has undergone a lot of effort, the preprocessing still has problems.

1.1 POS Tagger:

Part-of-speech (POS) employs tagging, which assigns the correct part of speech to each word in a phrase. It is useful to understand the grammatical structure of sentences by assigning each word a tag. It is the preprocessing task in natural language processing but still issues are present in POS tagging. The main issue is the same word has different POS tags occurring after processing. The POS tag includes the following words and their subtypes: nouns, verbs, adverbs, adjectives, pronouns, and conjunction. POS taggers allow automatic text processing tools to determine which part of speech each word belongs to, which is critical in pre-processing tasks before performing syntactic parsing or semantic analysis. In this paper, we try to focus on that issues and compare the result based on two POS tagging Tools i.e. NLTK and spaCy.

1.1.1 *NLTK:*

The NLTK is known as the Natural Language Toolkit which is a group of programming languages that uses symbols and statistics for English natural language processing (NLP). Steven Bird and Edward Loper designed it at the University of Pennsylvania's in Computer and Information Science Department. The NLTK provides wrappers for effective NLP commercial applications as well as language processing tools for classification, stemming, tokenization, tagging, parsing, and semantic reasoning. The NLTK library is great for working with natural language and a useful tool for teaching and working in computational linguistics with Python. For POS tagging, NLTK supports various languages. The off-the-shelf tagger is used for POS tagging in NLTK for English. This tagger used the Penn Treebank tagset for assigning the word [(12)].

1.1.2 *SpaCy:*

Python Natural Language Processing module is called spaCy and is open-source, free, and has many built-in features. Matthew and Ines, are the library's primary developers. It is distributed under the MIT license. Statistical models trained by well-known machine learning libraries can be connected using deep learning procedures supported by spaCy. For Named Entity Recognition (NER), dependency parsing, text categorization, and part-of-speech tagging, spaCy use convolutional neural network models. Word tokenization and POS tagging are performed in spaCy (11).

In this paper, we used the above POS tagging tool for comparison between the TREC (Text REtrieval Conference) data and data collected from the student. The POS tagger assigns a different tag for the same word and generates the incorrect result. So we have done a comparison between Standard data i.e.TREC Data and data gathered from students' data i.e. DS (Data Structure) Data. In the data, the question sentences and answer sentences are tagged with the POS tag for each word using both the libraries that are NLTK and spaCy. The comparison was done by calculating the dissimilarities in tags from both data.

2 Literature Survey

Four different part-of-speech taggers were used in this work (5) to show the core techniques (POS tagger). The ability to read the word's tag in the Question Answering field is a critical responsibility for the POS tagger. These include NLTK, FREELING, NLP tagger, and Cognitive POS tagger. They concentrate on the four POS tagging tools. Several differences for the same words have been noted. This is the category under which the same word is classified in many ways for parts of speech like NN, VB, JJ, RB, etc.

The information extraction method is made available for questions regarding natural language processing. In this study (6), the question-answering research has been thoroughly reviewed to address the problems caused by the explosion of information in this era of information and communication technology. The system also uses this method to obtain data depending on the issue. The likelihood that a potential response will be present and the words or non-letter characters that are usually connected with it will be used to assess whether it will match the question, even if some of the components of the question are missing. Using strategies for responding that are verbal, statistical, and pattern-based.

The current state (7) of the question-and-answer literature was summarized. They observed the paradigms and use them equally in their strategies, however, the hybrid technique has not been used consistently due to their complexity. Given that there aren't any implementations of question-answering systems in Brazilian Portuguese in the literature, the work's contributions comprise a thorough review of the research on these topics.

In this Paper (8) the result obtained here are very promising and can be improved by many actions including expanding the training corpus and improving the lexical analysis program. This is because they worked with a large and rich tagset and there isn't a standard truth corpus available. In order to maximize tagging accuracy, future work on this project would concentrate on increasing the size of the labeled corpus.

This review paper (9) offers an in-depth analysis of POS tagging methods using machine learning and deep learning techniques so that researchers can stay current on the field's developments. As a result of this review, it is clear that using deep learning techniques improves effectiveness and efficiency.

The majority of methods are data-based, where taggers get the proper way to classify unlabeled data from a corpus of pre-annotated data. Here is a succinct, up-to-date overview of POS tagging. POS tagging methods train computational trained models using tagged corpora.

This article (10) introduces three common types of tagging: statistical, rule-based, and evolutionary. Each common tagging method's benefits and drawbacks are presented and examined. 93–96% accuracy levels have been successfully attained for some rule-based and stochastic approaches, and 96–97% for some evolution algorithms.

The authors (11) e^{x}amined five well-known NER tools (Stanford NLP, NLTK, OpenNLP, SpaCy, and Gate) and presented the most recent versions of each in this paper. According to our reproducible approach, the results demonstrate that StanfordNLP outperforms the other evaluated applications on a selected corpus by 15% to 30%.For each evaluated piece of software, we were unable to find the same outcomes as those found in the literature. The observed difference may be as high as 66%.

The paper (12) Collected and researched numerous automated techniques used by papers for question generation while designing the system. Before creating a question, text that was provided as input is preprocessed. Semantic analysis and syntactic analysis are both included in per-processing. POS tagging and Chunking are a part of syntactic analysis. Semantic Analysis (SA) performs Named Entity Recognition (NER).

3 Proposed method

The proposed methodology will find out the dissimilarities that are present in the POS tagger for the same word using NLTK and spaCy library, for that we will compare the standard data i.e.TREC data and data collected from students that are denoted by DS Data. The dissimilarities are present in the word when the POS tagger is applied to the data and the same term is assigned a different POS tag. Therefore, using both libraries, we selected four Question words i.e. What, How, which, Name from both the data sets for questions and answers. We will focus on issues that are present in Preprocessing i.e. POS Tagger and analysis the result based on both the data.

Below Figure 1 shows the proposed method for the comparison of POS Tagging using in both libraries using NLTK and spaCy.

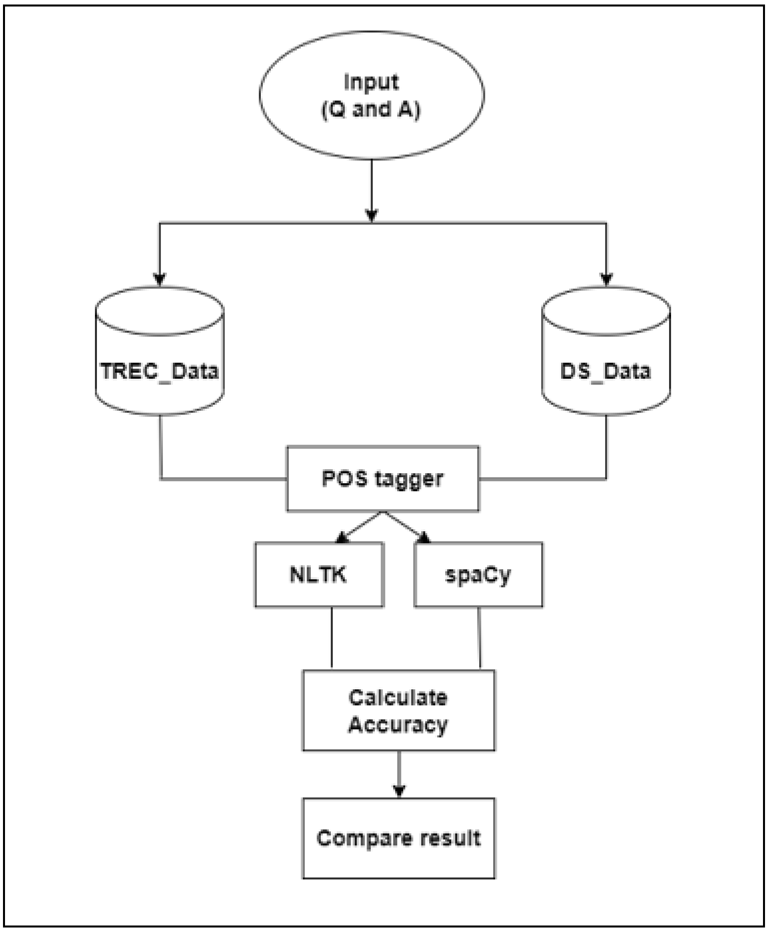


FIGURE 1

Proposed Method of POS Tagging

3.1 Input:

The input contains the question and answers that are taken from both data sets.

3.2 TREC_DATA:

The TREC is known as the Text Retrieval Conference is sponsored by the NIST Defense Department of the US. In this paper, TREC 2001 data set is used. In this data set total of 314 questions and answers are present.

Below is the sample of the TREC Data set Question and Answer:

```
QUESTION 894: How far is it from Denver to Aspen?
DOCNO: SJMN91-06154228
ANSWER: The Aspen/Snowmass area is about 200 miles southwest of Denver.

QUESTION 895: What county is Modesto, California in?
DOCNO: LA072190-0069
ANSWER: Modesto in Stanislaus County.

QUESTION 896: Who was Galileo?
DOCNO: FT924-10243
ANSWER: Galileo Galilei, the astronomer
```

FIGURE 2

Sample of TREC Question and Answer

3.3 DS_DATA:

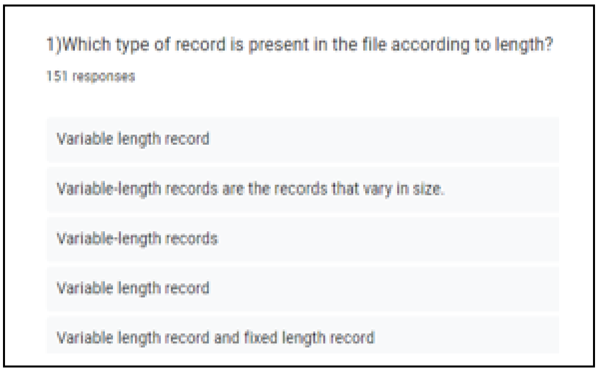
In this dataset, the teacher creates the questions, and answers are collected from students. The Questions were generated from the data structure subject in computer science. We created 20 Questions and got 151 responses from students that used them in our experiment.

Below is the Sample of Questions and Answer collected from students:

We calculate the POS tagger for TREC data and DS data and compare the result based on dissimilarities and calculate the accuracy.

4 RESULT AND ANALYSIS:

The result is obtained by comparing the two datasets produced by the POS tagger with the four question types we used: What, How, Name, and which. The difference between the libraries



1)Which type of record is present in the file according to length?
151 responses

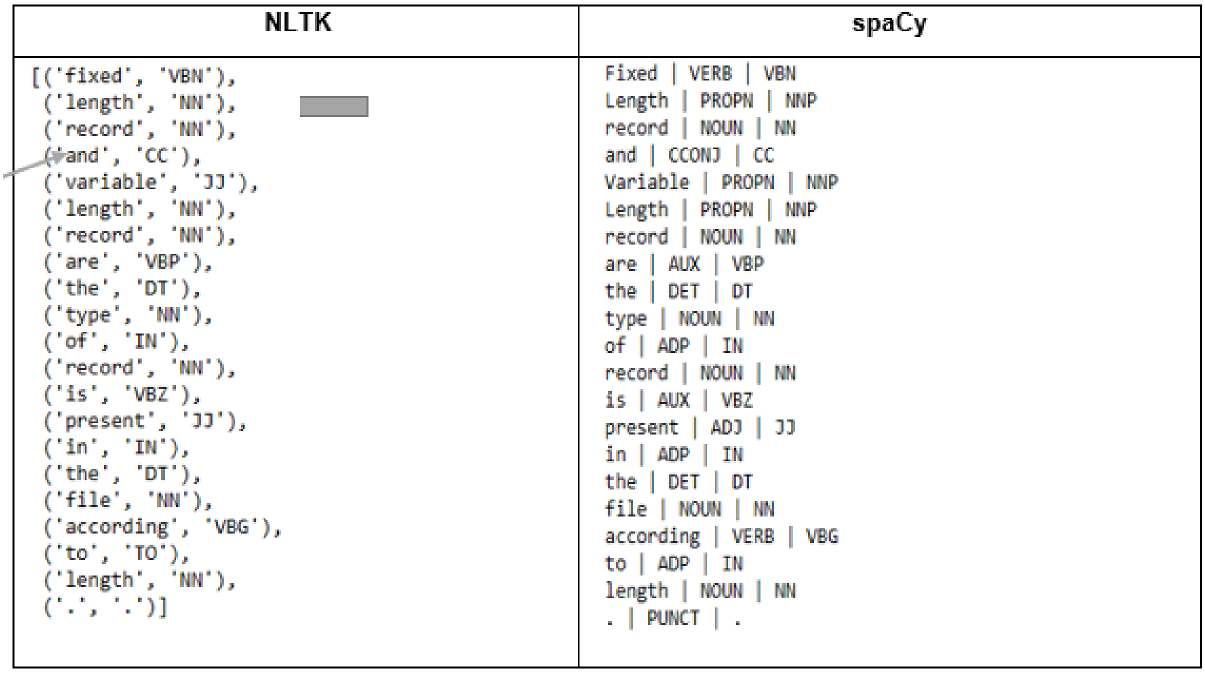
Variable length record
Variable-length records are the records that vary in size.
Variable-length records
Variable length record
Variable length record and fixed length record

FIGURE 3

Sample of responses collected from students

in the same word that includes various POS tags is seen after the POS tagger is performed. Using NLTK and spaCy, the same words generate different POS tags. The dictionary tag set was used to recheck those words, determine which library provided the correct response, and increase their count. Using the NLTK and spaCy libraries to compare the word with the dictionary tag and find the dissimilarities for the word.

Figure 4 showing the comparison of similarities and dissimilates of POS tagger for the same word after using both tool.



NLTK	spaCy
[('fixed', 'VBN'),	Fixed | VERB | VBN
('length', 'NN'),	Length | PROPN | NNP
('record', 'NN'),	record | NOUN | NN
('and', 'CC'),	and | CCONJ | CC
('variable', 'JJ'),	Variable | PROPN | NNP
('length', 'NN'),	Length | PROPN | NNP
('record', 'NN'),	record | NOUN | NN
('are', 'VBP'),	are | AUX | VBP
('the', 'DT'),	the | DET | DT
('type', 'NN'),	type | NOUN | NN
('of', 'IN'),	of | ADP | IN
('record', 'NN'),	record | NOUN | NN
('is', 'VBZ'),	is | AUX | VBZ
('present', 'JJ'),	present | ADJ | JJ
('in', 'IN'),	in | ADP | IN
('the', 'DT'),	the | DET | DT
('file', 'NN'),	file | NOUN | NN
('according', 'VBG'),	according | VERB | VBG
('to', 'TO'),	to | ADP | IN
('length', 'NN'),	length | NOUN | NN
('.', '.')]	. | PUNCT | .

FIGURE 4

comparison of similaritiesand dissimilates of POS tagger for the same word after using both tool

In above sample arrow indicate that the same word generate the different tag by applying NLTK and spaCy library. After finding we calculate the total correct count tag of question words for question and answer from the both of the data.

The below table1-4 displays the total right tag of the questions and responses from both sets of data.

TABLE 1

Total count of the correct tags of the questions and Answers with dictionary tags in what table

	What Table							
POS	**Question**				**Answer**			
Tag	**TREC_Data**		**DS_Data**		**TREC_Data**		**DS_Data**	
	NLTK	**spaCy**	**NLTK**	**spaCy**	**NLTK**	**spaCy**	**NLTK**	**spaCy**
NN	9	25	0	0	26	24	0	1
NNS	0	4			11	8	0	0
NNPs	0	3			0	1	0	2
NNP	3	4			3	19	0	0
VB	0	9			0	5	0	0
VBN	2	2			2	10	0	0
VBP	0	2			0	1	0	0
VBG	1	0			0	3	1	0
VBZ	0	0			0	7	0	1
VBD	0	0			0	0	1	0
JJS	0	1			1	0	0	0
RB	0	1			1	3	1	0
JJ	2	14			13	19	1	0
DT	1	0			1	0	0	0
UH	0	1			0	0	0	0
WDT	0	23			0	0	0	0

The above table shows the word with its Part Of Speech tag, which determines the total number of accurate tags in the NLTK and spaCy libraries. The accuracy of both libraries for each question type is then determined after the comparison is complete. The below formula is used to calculate the accuracy of both libraries (17) :

$$Accuracy = \frac{Total\ Words\ in\ NLTK\ or\ spaCy}{Total\ Word}$$

TABLE 2

Total count of the correct tags of the questions and Answers with dictionary tags in How table

POS Tag	How Table Questions TREC_Data NLTK	spaCy	DS_Data NLTK	spaCy	Answers TREC_Data NLTK	spaCy	DS_Data NLTK	spaCy
NN	0	1	0	1	0	0	0	1
NNP	0	2	0	0	0	0	0	0
NNS	0	0	0	2	0	1	1	0
NNPs	0	1	0	0	0	1	0	0
VB	0	1	0	0	0	0	0	0
VBN	1	1	1	0	0	0	0	0
VBP	0	1	0	1	0	0	0	0
VBG	0	0	0	1	0	0	0	0
VBZ	0	0	0	0	0	3	0	0
JJ	0	4	1	0	0	0	1	0
IN	0	1	0	0	0	0	0	0
RB	0	0	0	0	1	5	0	0

TABLE 3

Total count of the correct tags of the questions and Answers with dictionary tags in Which table

<table>
<tr><th></th><th colspan="4">Which Table</th><th colspan="4"></th></tr>
<tr><th>POS</th><th colspan="4">Questions</th><th colspan="4">Answers</th></tr>
<tr><th>Tag</th><th colspan="2">TREC_Data</th><th colspan="2">DS_Data</th><th colspan="2">TREC_Data</th><th colspan="2">DS_Data</th></tr>
<tr><th></th><th>NLTK</th><th>spaCy</th><th>NLTK</th><th>spaCy</th><th>NLTK</th><th>spaCy</th><th>NLTK</th><th>spaCy</th></tr>
<tr><td>NN</td><td>0</td><td>0</td><td rowspan="6">0</td><td rowspan="6">0</td><td>1</td><td>1</td><td rowspan="7">0</td><td rowspan="7">0</td></tr>
<tr><td>NNS</td><td>1</td><td>0</td><td>0</td><td>0</td></tr>
<tr><td>NNP</td><td>0</td><td>0</td><td>1</td><td>0</td></tr>
<tr><td>VBN</td><td>0</td><td>0</td><td>0</td><td>1</td></tr>
<tr><td>JJ</td><td>0</td><td>1</td><td>0</td><td>0</td></tr>
<tr><td>JJS</td><td>0</td><td>1</td><td>0</td><td>0</td></tr>
<tr><td>WDT</td><td>0</td><td>4</td><td>0</td><td>1</td><td>0</td><td>0</td></tr>
</table>

TABLE 4

Total count of the correct tags of the questions and Answers with dictionary tags in name table

POS Tag	Name Table							
	Questions				**Answers**			
	TREC_Data		**DS_Data**		**TREC_Data**		**DS_Data**	
	NLTK	**spaCy**	**NLTK**	**spaCy**	**NLTK**	**spaCy**	**NLTK**	**spaCy**
NN	3	0	1	0	0	3		
NNS	0	0	0	0	2	0	0	0
JJ	1	0	0	0	0	0		

TABLE 5

Accuracy in the percentage after comparing results for four question words

Q. word	Question in %				Answer in %			
	TREC_Data		**DS_Data**		**TREC_Data**		**DS_Data**	
	NLTK	**spaCy**	**NLTK**	**spaCy**	**NLTK**	**spaCy**	**NLTK**	**spaCy**
What	17	83	0	0	38	62	50	50
How	8	92	29	71	9	91	67	33
Which	14	86	0	10	50	50	0	0
Name	100	0	100	0	40	60	33	67

The accuracy of each question word in both datasets is shown in the table 5. The accuracy of both the libraries, NLTK, and spaCy, is individually determined using question-and-answer data for only four questions words. The analysis shows that the result of spaCy is good but the NLTK library generates the best correct tag for the 'How' Question word for answers. For Name, the NLTK is good for Question and for another Question word, spaCy generates good results.

5 Conclusion

To extract the right responses from a large amount of data, the Question Answering system is used. It is necessary to perform Preprocessing if we want the correct result for the data. We used a variety of methods for Preprocessing, but this paper focuses on the POS Tagger. In Python, different POS tagger libraries are present but sometimes they generate different results for the same word still this problem is present in the POS tagger. So we examine the differences between the tags using NLTK and spaCy libraries. In this paper, we used four question words i.e. What, How, Which, and Name that used to check the difference between the question data and answer data. We used two data for comparison, the first data is TREC data and the Second data is DS data. The comparative analysis was done on both the data using the libraries and then calculated the accuracy for question words and answer words in NLTK and spaCy. Overall, the accuracy of NLTK is 34.71% correct, whereas spaCy is 65.25% correct for questions using TREC DATA. The NLTK is 42.85% correct, whereas spaCy is 57.14% correct for questions using DS DATA. For Answers, NLTK is 34.18% correct, whereas spaCy is 65.81% correct for answers using TREC DATA. The NLTK is 50% correct, whereas spaCy is 50% correct for questions using DS DATA. Based on the above results, we find that spaCy is the best preprocessing tool for POS tagger than the NLTK. For preprocessing the researcher will use their own POS tagging rule with spaCy then which is good for their research.

Acknowledgment

The authors would like to acknowledge Chattrapati Shahu Maharaj Research Training and Human Development Institute (SARTHI), Pune for awarding a fellowship, also thanks to the Computational and Psycholinguistic Research Lab Facility for supporting this work and the Department of Computer Science and Information Technology, Dr. Babasaheb Ambedkar Marathwada University, Chhatrapati Sambhajinagar(Aurangabad), Maharashtra, India.

References

[1] T Brants, Natural Language Processing in Information Retrieval, 2003.

[2] M A Kankhar and C N Mahender, Online Subjective Examination a Student Perspective.

[3] M V S Rishita, M A Raju, and T A Harris. Machine translation using natural language processing. *MATEC Web of Conferences*, 277:2004–2004, 2019.

[4] S Adhikari. Nlp based machine learning approaches for text summarization. *2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC)*, pages 535–538, 2020.

[5] K B Khandale, A Pundge, and C N Mahender. Similarities in words Using Different Pos Taggers. *IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN*, pages 2278–0661.

[6] K Priya and T Harikrishnan, An Approach Of Information Extraction For Question Answering In Natural Language Processing.

[7] M A C Soares and F S Parreiras. A literature review on question answering techniques, paradigms and systems. *Journal of King Saud University-Computer and Information Sciences*, 32(6):635–646, 2020.

[8] D Kumawat and V Jain. POS tagging approaches: A comparison. *International Journal of Computer Applications*, (6):118–118, 2015.

[9] A Chiche and B Yitagesu. Part of speech tagging: a systematic review of deep learning and machine learning approaches. *Journal of Big Data*, 9(1):1–25, 2022.

[10] C Lv, H Liu, Y Dong, and Y Chen. Corpus based part-of-speech tagging. *International journal of speech technology*, 19(3):647–654, 2016.

[11] X Schmitt, S Kubler, J Robert, M Papadakis, and Y Letraon. A replicable comparison study of NER software: StanfordNLP, NLTK, OpenNLP, SpaCy, Gate. *Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, pages 338–343, 2019.

[12] P Gumaste, S Joshi, S Khadpekar, and S Mali. AUTOMATED QUESTION GENERATOR SYSTEM USING NLP LIBRARIES. *International Research Journal of Engineering and Technology (IRJET)*, 7(6):4568–4572, 2020.

[13] A Pinto, H Gonçalo Oliveira, and A Oliveira Alves. Comparing the performance of different NLP toolkits in formal and social media text. *5th Symposium on Languages, Applications and Technologies (SLATE'16). Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik*, 2016.

[14] R Jiang, R E Banchs, and H Li. Evaluating and combining name entity recognition systems. *Proceedings of the Sixth Named Entity Workshop*, pages 21–27, 2016.

[15] C Antoniou and N Bassiliades. A survey on semantic question answering systems. *The Knowledge Engineering Review*, pages 37–37, 2022.

[16] P J Antony and K P Soman. Machine transliteration for indian languages: A literature survey. *International Journal of Scientific & Engineering Research*, 2:1–8, 2011.

[17] X Schmitt, S Kubler, J Robert, M Papadakis, and Y Letraon. A replicable comparison study of NER software: StanfordNLP, NLTK, OpenNLP, SpaCy, Gate. *Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, pages 338–343, 2019.

[18] https://hasty.ai/docs/mpwiki/metrics/accuracy