Part-of-Speech Tagging accuracy for manufacturing process documents and knowledge

Fatemeh Besharati Moghaddam[1,2], Angel J. Lopez[1,2], Stijn De Vuyst[1,2], and Sidharta Gautama[1,2]

[1] Department of Industrial Systems Engineering and Product Design, Ghent University, Ghent, Belgium
[2] FlandersMake@UGent, corelab ISyE

Abstract Adaptive guidance systems in manufacturing that support operators during the assembly process need to serve the right information at the right time. A conversational recommender system as the single point of contact between the operator and different sources of information, based on natural language processing, can be introduced to assist the operators. Natural language processing techniques can help to mine answers in text-based knowledge repositories as available in training documents, work instructions, and company procedures. Both the content as well as the style of writing in these documents are different from general language use and we examine the accuracy of part-of-speech tagging within this close domain of manufacturing. A benchmark dataset has been constructed based on four different classes of documents typical in the manufacturing domain. The dataset contains 1206 tokens divided over eight tag types. The accuracy of two open-source corpora, spaCy and NLTK, has been measured on this benchmark with an average accuracy of resp. 93% and 87%. The conclusion drawn is that pre-trained natural language libraries can effectively handle the specific contexts in the assembly domain based on the provided accuracy.

Keywords: Natural Language Processing, Part of Speech, Closed domain, Operator support

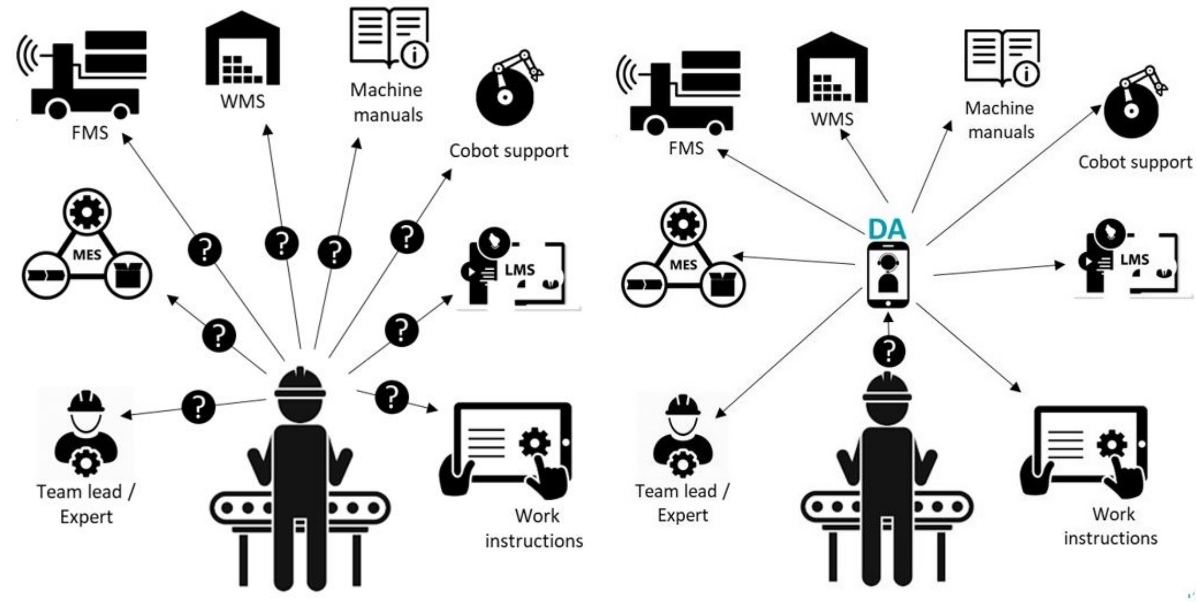
1 Introduction

In recent years, manufacturing companies have faced an increasing demand for complex products with more variations, leading to assembly operators needing to master more tasks. The introduction of Industry 4.0 technologies has also meant that operators need to acquire new skills to handle the complexity of different products. Prinz et al. [24] introduce the operator assistance systems as one of the branches of research related to Industry 4.0.

Currently, support services for assembly operators typically only provide procedural digital work instructions, leaving operators to rely on their own knowledge or seek help from external experts in case of issues. Digitization presents an opportunity to gather information from IT systems and provide better support to operators [34]. There are different developments regarding adaptive assistance systems in manufacturing contexts, e.g., Head-Mounted Displays [15], Augmented Reality [10, 27], Tangible User Interfaces [26], and Motion Recognition [31].

Nonetheless, all of the support systems referred to earlier are pre-programmed and have well-organized information. The difficulty that arises in the assembly area may be linked to unstructured open questions posed by the operator. In this situation, the assistance system ought to analyze the open question and search various knowledge databases to locate appropriate and effective responses. Figure 1a shows the current situation of operator assistance in the assembly domain. For the operator to have efficient access to the needed information, there is a need for a direct interactive portal that interacts with the operator and serves as a single point of contact between the operator and various information sources [21].

Thus, the conversational recommender system as a digital interface based on natural language understanding aims to assist users in retrieving or suggesting the most relevant information by means of textual or spoken dialog. This enables users to interact with the system more effectively by employing natural language processing (NLP) [37]. Having this digital assistance can be key in increasing the operator assistance efficiency in the assembly line in order to handle the various open questions (Figure 1b).



(a) Current situation, without access to the digital assist system (b) Proposed situation, with access to the digital assist system

Figure 1: In the first figure, the operator has access to different knowledge databases to use in the problematic situation, but access is unrecognized, fragmented, and therefore inefficient. However, in the second one, with the help of a digital assist operator can faster and easier access the needed information

Natural Language Processing [32] is helpful in understanding open questions from the operator in the assembly domain. This involves matching the query with an existing list of issues and searching for potential solutions/answers across various knowledge databases. NLP is the technique to help the computer and machines process in order to understand the contents and texts from human language. In other words, NLP is helping to improve communication between humans and machines [32].

In the NLP area, research is already done on the open domain and the closed domain [18]. Open domain and closed domain are terms used in NLP to describe the scope of a given language model's knowledge and the range of questions or tasks it can effectively handle. The open domain language model is designed to handle a broad range of topics and questions [35]. There are many types of research in the open domain of NLP with efficient results [30].

On the other hand, the closed domain language model is designed to perform specific tasks within a restricted area or field [4]. To the best of our knowledge, there is limited existing research for implementing NLP techniques in the manufacturing area, especially in the assembly domain. Shi [28] and Xingguang et al. [33] propose the question-answering system for the manufacturing domain. However, the details about the domain are not mentioned in their research. We assume that they considered the open corpus for their approach. So, there is a gap in the state-of-the-art for checking the NLP techniques in the assembly considering the close domain corpus. To start work in the mentioned gap, we develop the part of speech tagging technique in NLP for the assembly domain.

The challenge for assembly as a specific domain lies in (1) a lack of publicly available documents and specific corpora related to NLP in assembly, (2) small document libraries within organizations, and (3) informal writing and poor grammar in procedures like quality reports. Though the first challenge of this research is to check, by preparing a benchmark data set in NLP related to the assembly domain if the existing pre-trained NLP libraries are capable of processing the various assembly-related concepts in the NLP domain.

In the following, section 2 reviews related work, Section 3 introduces the problem and the methodology used in this research, Section 4 presents the experiment results, and the conclusion and future works are presented in section 5.

2 Literature Review

One of the most important areas in the pre-processing steps of NLP is Part-of-Speech (POS) tagging. POS tagging is the task of labeling or tagging each token in sentences based on the defined rule [12, 6]. POS tagging is useful for a variety of NLP tasks, such as information extraction, entity recognition, and grammatical structure identification. It automatically assigns the parts of speech tags to the tokens considering two main aspects: finding the exact tags for each token and choosing between the possible tags for ambiguous tokens [14, 29, 8].

The main goal of developing the POS tagger for any language is improving the accuracy of tagging and also considering the different language structures, trying to remove the ambiguity in the tokens [7]. Based on [16], the number of tokens in the training and testing data and also the corpus or the open-source dictionary being used in POS tagging can be two important factors in the performance and accuracy of POS tagging.

In some research, authors used the manually trained corpus in the open domain [1, 3] or the closed domain [20, 36] for their experiment. Kumar [11] proposed an approach in POS tagging considering their defined corpus with 77860 tokens for training and 7544 for testing. In [19], 14369 tokens in the training set and 5000 tokens in the testing set are studied. Rezai [25] offered a POS tagger corpus with 5000000 tokens for training and 11766 tokens in the test set for the Persian language.

However, using the manually annotated corpus, the corpus size may not be enough for modeling and an efficient evaluation [5, 22]. Open-source NLP libraries can be used by authors to train their methodology and test their data set, such as the Stanford NLP suite [17], Google SyntaxNet [23], NLTK [13], and SpaCy [9]. [2] used four different open-source libraries in order to train their methodology individually. In the next step, they manually annotated 1116 tokens with the correct part-of-speech tag and test the tagging accuracy based on each open-source library.

Based on the state-of-the-art, there is a gap in available specific corpora related to NLP in the assembly domain. In this work, we try to address this gap with the help of two pre-trained available libraries. we aim to identify how the choice of using a particular open-source available NLP library could impact the results of POS tagging in the assembly domain.

3 Methodology

POS tagging is a crucial step in NLP in the case of improving the performance of a system related to information retrieval [16]. Based on [16], we consider eight classes of parts of speech tags as Noun, Verb, Adjective, Pronoun, Determiner, Adverb, Preposition, and Conjunction.

In order to implement POS tagging, there is a wide variety of NLP libraries used in research related to natural language tagging. In this research, we used two open-source NLP libraries SpaCy and NLTK (Natural Language Toolkit). NLTK is a leading platform for building Python programs to work with human language data. SpaCy is a library for advanced NLP written in Python and Cython.

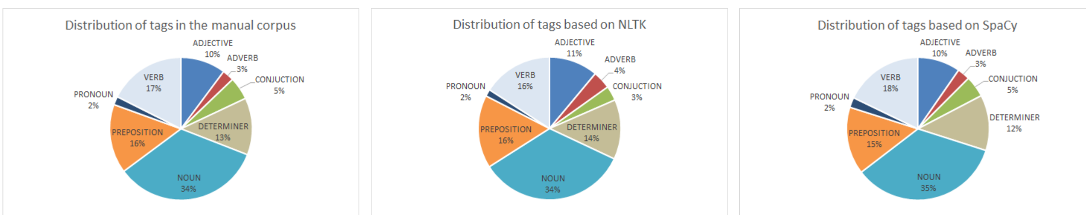
To the best of our knowledge, there are no publicly available POS annotated training data on the assembly domain of manufacturing. Thus, we create a set including 100 sentences picked from four different categories in the assembly domain, Warning, Informative-texts, Manual, and Work-instructions. These four different categories are chosen based on the structure of texts and sentences. In assembly, some of the used documents are well-written structured sentences formulated in the PDF format called manual documents. A sample of these manual sentences can be "The size and load-carrying capacity of the groove increases in line with the modular dimension". In the prepared benchmark, we consider around half of the sentences to the well-written structured manual sentences.

In addition, in the assembly line, there are some semi-structured documents like the warning and informative texts and the work instruction detail for each activity. These types of texts are usually written as imperative sentences or non-completed sentences. The samples can be "Assemble three profiles with angle brackets." or "Part extremely hot!!!". In the prepared benchmark, we consider the other half of the sentences as three different sub-categories in semi-structured sentences (Warning, Informative-texts, and Work-instructions).

Having 1206 tokens (excluding the punctuation marks), we do the annotation for all the tokens manually. The introduced eight POS tags based on our reference [16] assigned to each token in 100 sentences. After doing the manual POS tagging with the help of an expert, we investigate which of the considered open-source libraries achieves the best result in the test set. After manual tag the considered corpus, we have 212 verbs, 23 pronouns, 189 prepositions, 409 nouns, 156 determiners, 62 conjunctions, 32 adverbs, and 123 adjectives in our annotated data set. Figure 2a, shows the distribution of the tags in the considered corpus based on the assembly context.

4 Result and Discussion

In order to evaluate if NLTK and SPACY pre-trained taggers libraries can perform well in assembly domain corpus, we use the ground truth consisting of 1206 tokens in 100 different category sentences, we implement the libraries taggers into the considered ground truth in order to compare the result of different tags with the manual annotation.



(a) Tags in ground truth (b) Tags in NLTK taggers (c) Tags in SPACY taggers

Figure 2: The distribution of tags based on 8 different categories for the considered assembly corpus with 1206 tokens in the ground truth, NLTK taggers, and SPACY taggers

As shown in Figure 2b, the distribution in the tag predictions for different tags with the help of the NLTK library is approximately near the manual ground truth. The tagging accuracy for the NLTK tagger is 87% compared to the ground truth. For the SpaCy library, the distribution of different tags has a high similarity with the manually annotated corpus (Figure 2c). The tagging accuracy for the SpaCy tagger on the ground truth is 93%, which in comparison to NLTK is 6% more accurate.

In order to recognize all the tokenized words in comparison to the ground truth, the NLTK tagger could identify 1206 tokens exactly the same as the annotated ground truth. So, identical tokens accuracy is 100% for NLTK tagger. For the hyphenated Compound words, the NLTK tagger could recognize them (e.g.,difficult-to-reach, self-threading) with the same structure as manual ground truth and consider them as one specific token. The SpaCy tagger identifies all the tokens in the corpus. However, the number of tokens identified with the SpaCy tagger is 1223. The reason is that for the hyphenated compound words, SpaCy tagger tokenizes the words separately. So, for example, a hyphenated compound word as *difficult-to-reach* in SpaCy tagger acknowledge as three separated tokens *difficult*, *to*, and *reach*. In the considered assembly corpus, we have 20 hyphenated compound tokens.

Figure 3 shows the accuracy of each individual tag for the considered libraries. The three most important parts of speech are nouns, verbs, and adjectives. SPACY achieves 92% for verbs, 94% for nouns, and 76% for adjectives, and for NLTK is this 81%, 89%, and 79%. SPACY performs better on the considered ground truth for nouns and verbs and in addition for pronouns. In recognizing adverbs, determiners, and prepositions, both libraries gained efficient accuracy. For conjunction tags, NLTK obtained 63% accuracy as the worse accuracy compared to other tags.

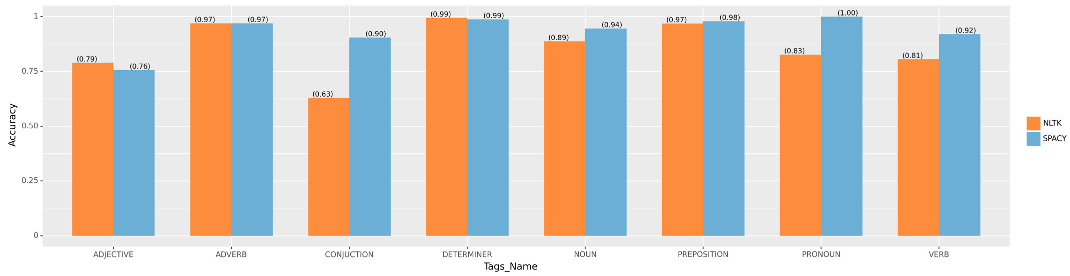


Figure 3: Accuracy for each individual tag on the considered assembly corpus of NLTK and SPACY libraries taggers

The miss-annotation accuracy probability for the tags as $P(PredTag|TrueTag)$ can be calculated for each tag in each library. This probability can be estimated simply by counting the number of occurrences of each token if it is annotated by the pre-trained library in a different tag compared to the ground truth. For example, $P(Verb|Noun)$ is the probability that a specific token that is in the ground truth annotated as a noun, would be annotated as a verb by the pre-trained tagger in the considered library. The lower value leads to better annotation accuracy. In Table 1, the highest miss-annotation probabilities are shown in percentage. In NLTK tagger, 29 verb tokens are miss-annotated as nouns. Based on the deeper analysis of the prediction results in NLTK, most of these miss-annotations are related to semi-structured sentences that start with a verb instead of a noun. However, in well-written texts e.g., manuals, 18 adjective tokens are miss-annotated as nouns which in most cases it is related to several nouns written continuously in a sentence. For conjunction tokens that have the most miss-annotation compared to other tokens, 13 out of 63 are classified as prepositions by NLTK tagger. The frequency of this is equal for well-written texts and semi-structured sentences. In the SPACY tagger, 12 verb tokens classify as nouns the same as the NLTK tagger mostly happening in imperative sentences. 19 adjective tokens are tagged as nouns and 12 adjective tokens as verbs. These miss-taggings happen in well-structured sentences when a series of nouns and adjectives are continuously written in a sentence.

Table 1: Probability of more frequent miss-annotated tags considering the assembly ground truth in NLTK and SPACY libraries

NLTK		SPACY	
$P(PredTag\|TrueTag)$	$(PredTag\|TrueTag)$	$P(PredTag\|TrueTag)$	$(PredTag\|TrueTag)$
$P(Noun\|Verb)$	29|211	$P(Noun\|Verb)$	12|211
$P(Noun\|Adjective)$	18|123	$P(Noun\|Adjective)$	19|132
$P(Preposition\|Conjunction)$	13|63	$P(Verb\|Adjective)$	12|132

5 Conclusion and Future work

In order to give efficient support to the operator in the assembly line, the digital assistant should be able to link the open question from the operator to the needed information with the help of NLP. Due to the lack of an efficient corpus in the assembly domain, we introduce a manual benchmark with 100 sentences related to four different categories in the assembly. Considering 1206 tokens in the prepared ground truth and with the help of an expert, we labeled each token based on POS tagging techniques. Using two NLP open-source libraries SPACY and NLTK, we checked if, with the help of their pre-trained taggers, we can achieve efficient accuracy in tagging the assembly concepts.

Our results show that the tagging accuracy for SPACY is 6% more accurate than NLTK with 93% accuracy. Both of the libraries recognized all the tokens in the ground truth. However, SPACY had some problems with tokenizing the hyphenated compound words as one dedicated token. In addition, the complete analysis is already done based on each individual tag for both libraries considering well-structured and also semi-structured sentences. Based on the result and efficient accuracy, we can conclude that the open-source libraries in NLP can have the potential to handle the specific concepts in the close domain of assembly considering the prepared benchmark.

Knowing that the open-source libraries in NLP efficiently can understand the assembly concepts, future work would focus on developing the conversational recommender system in order to propose efficient support to the operator in regard to the input question coming from the operator.

Bibliography

[1] K. Akhil, R. Rajimol, and V. Anoop. Parts-of-speech tagging for malayalam using deep learning techniques. *International Journal of Information Technology*, 12:741–748, 2020.

[2] F. N. A. Al Omran and C. Treude. Choosing an nlp library for analyzing software documentation: a systematic literature review and a series of experiments. In *2017 IEEE/ACM 14th international conference on mining software repositories (MSR)*, pages 187–197. IEEE, 2017.

[3] D. Anastasyev, I. Gusev, and E. Indenbom. Improving part-of-speech tagging via multi-task learning and character-level word representations. *arXiv preprint arXiv:1807.00818*, 2018.

[4] A. Arbaaeen and A. Shah. Ontology-based approach to semantically enhanced question answering for closed domain: A review. *Information*, 12(5):200, 2021.

[5] 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.

[6] C. Chungku, J. Rabgay, and G. Faaß. Building nlp resources for dzongkha: a tagset and a tagged corpus. In *Proceedings of the Eighth Workshop on Asian Language Resouces*, pages 103–110, 2010.

[7] D. L. Cing and K. M. Soe. Improving accuracy of part-of-speech (pos) tagging using hidden markov model and morphological analysis for myanmar language. *International Journal of Electrical and Computer Engineering*, 10(2):2023, 2020.

[8] B. R. Das, S. Sahoo, C. S. Panda, and S. Patnaik. Part of speech tagging in odia using support vector machine. *Procedia Computer Science*, 48:507–512, 2015.

[9] A. Explosion. spacy-industrial-strength natural language processing in python. *URL: https://spacy.io*, 2017.

[10] D. Gorecky, R. Campos, H. Chakravarthy, R. Dabelow, J. Schlick, and D. Zühlke. Mastering mass customization–a concept for advanced, human-centered assembly. *Academic Journal of Manufacturing Engineering*, 11(2), 2013.

[11] S. Kumar, M. A. Kumar, and K. Soman. Deep learning based part-of-speech tagging for malayalam twitter data (special issue: deep learning techniques for natural language processing). *Journal of Intelligent Systems*, 28(3):423–435, 2019.

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

[13] E. Loper and S. Bird. Nltk: The natural language toolkit. *arXiv preprint cs/0205028*, 2002.

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

[15] C. Magerkurth, T. Engelke, and C. Röcker. The smart dice cup: a radio controlled sentient interaction device. In *Entertainment Computing-ICEC 2006: 5th International Conference, Cambridge, UK, September 20-22, 2006. Proceedings 5*, pages 211–216. Springer, 2006.

[16] C. Manning and H. Schutze. *Foundations of statistical natural language processing*. MIT press, 1999.

[17] C. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. Bethard, and D. McClosky. Stanford corenlp a suite of core nlp tools, 2014.

[18] A. Mishra and S. K. Jain. A survey on question answering systems with classification. *Journal of King Saud University-Computer and Information Sciences*, 28(3):345–361, 2016.

[19] S. Mohammed. Using machine learning to build pos tagger for under-resourced language: the case of somali. *International Journal of Information Technology*, 12(3):717–729, 2020.

[20] E. Mutabazi, J. Ni, G. Tang, and W. Cao. A review on medical textual question answering systems based on deep learning approaches. *Applied Sciences*, 11(12):5456, 2021.

[21] D.-M. Park, S.-S. Jeong, and Y.-S. Seo. Systematic review on chatbot techniques and applications. *Journal of Information Processing Systems*, 18(1):26–47, 2022.

[22] A. H. Patoary, M. J. B. Kibria, and A. Kaium. Implementation of automated bengali parts of speech tagger: an approach using deep learning algorithm. In *2020 IEEE Region 10 Symposium (TENSYMP)*, pages 308–311. IEEE, 2020.

[23] S. Petrov. Announcing syntaxnet: The world's most accurate parser goes open source. *Google Research Blog*, 12, 2016.
[24] C. Prinz, D. Kreimeier, and B. Kuhlenkötter. Implementation of a learning environment for an industrie 4.0 assistance system to improve the overall equipment effectiveness. *Procedia manufacturing*, 9:159–166, 2017.
[25] M. J. Rezai and T. Mosavi Miangah. Farsitag: A part-of-speech tagging system for persian. *Digital Scholarship in the Humanities*, 32(3):632–642, 2017.
[26] C. Röcker. Universal access to awareness information: using smart artefacts to mediate awareness in distributed teams. *Universal Access in the Information Society*, 11:259–271, 2012.
[27] C. Röcker and R. Etter. Social radio: A music-based approach to emotional awareness mediation. In *Proceedings of the 12th international conference on Intelligent user interfaces*, pages 286–289, 2007.
[28] M. Shi. Knowledge graph question and answer system for mechanical intelligent manufacturing based on deep learning. *Mathematical Problems in Engineering*, 2021:1–8, 2021.
[29] J. Singh, N. Joshi, and I. Mathur. Part of speech tagging of marathi text using trigram method. *arXiv preprint arXiv:1307.4299*, 2013.
[30] 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.
[31] N. Ukita, D. Kaulen, and C. Röcker. A user-centered design approach to physical motion coaching systems for pervasive health. *Smart Health: Open Problems and Future Challenges*, pages 189–208, 2015.
[32] S. Vajjala, B. Majumder, A. Gupta, and H. Surana. *Practical natural language processing: a comprehensive guide to building real-world NLP systems*. O'Reilly Media, 2020.
[33] L. Xingguang, C. Zhenbo, S. Zhengyuan, Z. Haoxin, M. Hangcheng, X. Xuesong, and X. Gang. Building a question answering system for the manufacturing domain. *IEEE Access*, 10:75816–75824, 2022.
[34] X. Yang and D. A. Plewe. Assistance systems in manufacturing: A systematic review. In *Advances in Ergonomics of Manufacturing: Managing the Enterprise of the Future: Proceedings of the AHFE 2016 International Conference on Human Aspects of Advanced Manufacturing, July 27-31, 2016, Walt Disney World®, Florida, USA*, pages 279–289. Springer, 2016.
[35] Z. Ye, Z. Jia, Y. Yang, J. Huang, and H. Yin. Research on open domain question answering system. In *Natural Language Processing and Chinese Computing: 4th CCF Conference, NLPCC 2015, Nanchang, China, October 9-13, 2015, Proceedings 4*, pages 527–540. Springer, 2015.
[36] S. Yitagesu, X. Zhang, Z. Feng, X. Li, and Z. Xing. Automatic part-of-speech tagging for security vulnerability descriptions. In *2021 IEEE/ACM 18th International Conference on Mining Software Repositories (MSR)*, pages 29–40. IEEE, 2021.
[37] Y. Zhang, X. Chen, Q. Ai, L. Yang, and W. B. Croft. Towards conversational search and recommendation: System ask, user respond. In *Proceedings of the 27th acm international conference on information and knowledge management*, pages 177–186, 2018.