DOI: 10.23977/langl.2024.070408
ISSN 2523-5869 Vol. 7 Num. 4

The Scope of Part of Speech Tagging: A Bibliometric Study

Xiaoyi Du[1,a,*], Litong Wu[2,b], Bingliang Zhang[2,c], Xinye Li[2,d], Feng Hu[3,e]

[1]Research Center for Language and Language Education, Central China Normal University, Wuhan, China
[2]College of Economics and Management, Zhejiang Normal University, Jinhua, China
[3]School of Business, Taizhou University, Taizhou, Jinhua, China
[a]duxy_1982@163.com, [b]jun807325@163.com, [c]zhangbingliang1228@163.com, [d]seeandknow@163.com, [e]hufeng_1978@126.com
[*]Corresponding author

Keywords: Part of Speech, Natural Language Processing, Bibliometric analysis, Research scope

Abstract: POS tagging is a process of identifying the part of speech of a word in a text by considering the context in which it appears. To better understand the intellectual framework of POS tagging research, we conducted a thorough analysis of POS literature available in the Web of Science repository. By using co-word, co-citation, and social network analysis techniques on 1,656 relevant articles and 69,357 cited references, we were able to identify the main research topics and research streams related to POS tagging. We have explained each of the research streams in detail, along with an informative visualization that shows the evolution of research streams over time and the intellectual structure. After that, we have provided a comprehensive discussion of the findings, highlighting the current hotspots and future prospects in POS tagging research.

1. Introduction

The computer process of natural language processing (NLP) involves a series of steps. Part-of-speech (POS) tagging normally is one of the earliest stages for language pre-processing and understanding. POS tagging tries to label a part of speech for each word in the context. From a practical point of view, POS tagging is employed to extract enough information about the grammatical behavior of a word. Taking the representative noun as an example, a noun can play as the head of a noun phrase or an object of a verb or adjective. While the noun and noun phrases consist of the underlying entities in some NLP tasks, such as topic detection. POS tagging, therefore, plays a key role in the first stages of most NLP projects.

In the past twenty years, researchers have shown great interest in the corpus and approaches of POS tagging. The prior efforts normally dealt with the rule-based labeling of POS in different language or application scenarios, while the latest research focused on the efficiency improvement of POS tagging. Manual ling is laborious and expensive, hence widespread interest transferred to automating the tagging process, such as the Hidden Markov model (HMM), Deep learning (DL), and long short-term memory (LSTM). The scope of POS tagging has partly been disclosed in previous literature [1, 2]. However, existing studies rarely systematically revealed the research paradigms and application usages of POS tagging. The current study attempts to employ bibliometric analysis to identify the research scope of POS tagging within a scientific field.

2. Methodology

We use the bibliometric analysis method to explore the intellectual structure of POS tagging research. The bibliometric method is defined as the application of quantitative tools to bibliographic data [3]. It enables researchers to handle large quantities of bibliographic data while simultaneously minimizing any potential biases. This paper adopts co-citation, which denotes a joint citation of two articles in a later article to fulfill the research objectives.

2.1. Data collecting and filtering

The literature on POS Tagging is collected from Social Science Citation Index and Science Citation Index Expanded (SSCI & SCIE) on August 16, 2022. To ensure the quality and understandability of the work dataset, only double-blind, peer-reviewed journal articles are reminded. The data includes all journal studies on POS Tagging between 2008 and 2022. In the current study, Bibexcel software is employed for initial data processing. We extracted citing articles and cited references from WOS data respectively, and built serial number indexes for all of them. For each cited reference, we set its digital object unique identifier (DOI) as ID for the following informatics analysis.

2.2. Detecting the research streams

By scrutinizing 921 article publications and 40,647 cited references based on co-citation analysis, the detailed implications of intellectual discourses of POS Tagging can be uncovered. This study employed exploratory factor analysis to identify research streams, representative and influential publications [4]. Factor analysis evaluates the weight of publication representativeness and influence by comparing its factor loadings and factor scores among different factor categories. Factor loading (FL) of an article indicates how well this article fits into a special factor, and the factor score (FS) of a paper presents its contribution weight to a special discourse [5].

2.3. Social network analysis and time-series analysis

Social network analysis is the most popular approach for mapping the structure of the entire research network and uncovering the knowledge exchange between discourses [6]. It provides a big picture to visualize the structure of overall research and show the communication within and between research streams. Time-series analysis refers to analyzing a single set of time-indexed observations in a univariate context to comprehend dynamic change. Since data is usually distributed over time (e.g., years), time series analysis can provide a holistic view of the research stream over different time periods, understand the context of historical research, and predict their future developing direction.

3. Key research streams

To detect the research streams, we conduct factor analysis on 221 top-cited references (cited frequency > 5, approximately equal to 0.5% of all cited references) of POS Tagging. In total, the six streams explain 68.2% of variance in the data. An overall MSA of 0.872 and a significant Bartlett's test indicate that factor analysis was appropriate for the data. The proportion of explained variance indicates the importance of a stream to the field's theoretical foundation [4]. Table 1 presents the top 10 most representative publications (highest FL) in each research discourse and their influence (FS) on its category.

Table 1: Top 10 representative publications in each discourse.

	1: State-of-the-art models			**2: Sentiment Classification**			**3: Aspect-based Sentiment**		
	(Variance explained: 21.2%)			(Variance explained: 15.5%)			(Variance explained: 10.6%)		
N	Publication	FL	FS	Publication	FL	FS	Publication	FL	FS
1	Ling et al. (2015)	0.85	1.60	Subrahmanian & Reforgiato (2008)	0.87	1.89	Fan et al. (2018)	0.87	2.70
2	Ma & Hovy (2016)	0.83	3.65	Taboada et al. (2011)	0.86	4.17	Xue & Li (2018)	0.79	2.16
3	Lample et al. (2016)	0.81	3.65	Xia et al. (2011)	0.85	4.06	Chen (2017)	0.78	4.52
4	Huang & Chen (2015	0.81	1.95	Blitzer et al. (2007)	0.85	2.00	Dong et al. (2014)	0.78	3.23
5	Rei (2017)	0.78	0.54	Liu (2012)	0.83	2.54	Pontiki et al. (2014)	0.78	3.24
6	Schuster & Paliwal (1997)	0.77	1.29	Tan & Zhang (2008)	0.83	2.63	Wang et al. (2016)	0.76	5.29
7	dos Santos & Gatti (2014)	0.77	1.72	Pang & Lee (2008)	0.83	4.22	Cambria (2016)	0.70	1.88
8	Bojanowski et al. (2017	0.75	2.24	Pang & Lee (2004)	0.80	2.69	Liu (2011)	0.67	2.20
9	Chiu & Nichols (2016)	0.74	0.62	Moraes et al. (2013)	0.80	1.67	Vaswani et al. (2017)	0.61	2.90
10	Srivastava et al. (2014)	0.73	2.12	Agarwal et al. (2015)	0.80	2.64	Qiu et al. (2011))	0.60	1.04
	4: Topic modelling			**5: Word Frequency**			**6: Methods of POS tagging**		
	(Variance explained: 7.4%)			(Variance explained: 7.1%)			(Variance explained: 6.4%)		
N	Publication	FL	FS	Publication	FL	FS	Publication	FL	FS
1	Deerwester et al. (1990)	0.67	1.20	Brysbaert et al. (2012)	0.91	3.75	Brill (1995)0.70	2.16	
2	Rumelhart & Hintont (2019)	0.65	1.83	Balota et al. (2007)	0.89	4.99	Xue et al. (2005)	0.67	2.35
3	Bengio et al. (2001)	0.64	7.27	Baayen et al. (2011)	0.87	2.35	Petrov et al. (2006)	0.63	0.86
4	Elman (1990)	0.58	1.69	Brysbaert & New (2009)	0.87	6.07	Toutanova et al. (2003)	0.56	6.62
5	Blei et al. (2003)	0.58	3.49	Cai and Brysbaert (2010)	0.85	4.71	P.~Brown et al. (1992)	0.55	3.70
6	Tomas (2010)	0.55	1.60	Keuleers et al. (2010)	0.85	4.71	Cohen (1960)	0.55	1.54
7	Chen et al. (2017)	0.54	0.66	Dimitropoulou et al. (2010)	0.85	4.71	Maamouri et al. (1995)	0.54	0.78
8	LeCun et al. (1998)	0.42	0.36	New et al. (2007)	0.85	4.71	Porter (1980)	0.49	0.47
9				PAIVIO et al. (1968)	0.53	0.48	Santorini (1990)	0.48	0.59

3.1. State-of-the-art models

The first research stream explains 21.2% of the total variance and deals with State-of-the-art models to improve the accuracy and/or efficiency of POS Tagging. To figure out the exploding/vanishing gradient problems when learning long-term dependencies, Hochreiter and Schmidhuder (1997) introduced a gradient-based method called long short-term memory (LSTM)

[7]. LSTM has impacted several practical and theoretical fields, e.g., both Google and Facebook applied it to improve machine translation systems. LSTM constitutes the underlying methodology of most state-of-the-art models involved in POS Tagging, so that Hochreiter and Schmidhuder's study contributes most (FL=0.69, FS=5.01) to this discourse. The most representative reference (FL=0.85) in this discourse is Ling (2015) [8], which introduced a novel word representation model based on bidirectional LSTM. Publications in this discourse commonly share the same research path: Develop/design State-of-the-art models based on classic models/algorithms.

3.2. Sentiment Classification

The second research stream explains 15.5% of the total variance and focuses on the use of sentiment classification based on POS Tagging. In general, sentiment classification consists of two approaches: the supervised method and unsupervised method. The former applies a machine learning algorithm for text auto-categorization based on a training sample. The key to supervised machine learning is the engineering of a set of effective features, such as part of speech, sentiment terms, and shifters. The latter performs subjectivity classification based on a lexicon dictionary composed of sentiment words or fixed syntactic patterns composed by POS tags. The most influential article (FL=0.76, FS=7.96) in this discourse is Pang (2002) [9], which examines the efficiency of machine learning techniques in subjectivity classifying of movie reviews and compares these methods with traditional topic-based categorization. Subrahmanian's (2008) research [10], which has the highest level of representativeness (FL=0.87), introduces a comprehensive framework that covers all adjectives, verbs, and adverbs to identify opinions on any given topic, which is an improvement over previous sentiment classification methods that only analyze a single part-of-speech.

3.3. Aspect-based Sentiment

The third research stream explains 10.6% of the total variance and deals with the development of aspect-based sentiment analysis. Sentiment analysis is a vital NLP task that received increasing attention in recent years. However, early sentiment analysis commonly focuses on assessing the overall subjectivity of a given text, ignoring the concerned entities and/or aspects. Aspect-based sentiment analysis (ABSA) is a novel fine-grained technique in sentiment classification, which provides sentiment polarities of a given aspect or entity in a text. Instead of assessing the overall subjectivity of a sentence/document, ABSA is developed to detect the sentiment polarity of a given entity or aspect category, thus it enables to better understand the writers' opinions on a fine-grained level. Wang (2016) contributes the most to this discourse by introducing an attention-based LSTM Network for ABSA [11]. The novel model is easy to train and can detect sentiment polarity of given aspects and entities efficiently. Fan (2018) is the top two representative paper developing a novel framework of ABSA based on state-of-the-art algorithms [12].

3.4. Topic Modelling

The fourth research stream explains 7.4% of the total variance and focuses on topic modelling based on POS Tagging. The difficulty of language modelling is the curse of dimensionality, Bengio's (2001) article [13], the most influential work (FS=7.27), introduces a neural probabilistic language model to figure out the problem, which addresses the representation of word distribution and sequence simultaneously in a state-of-the-art trigram model. The novel model works based on using prior linguistic knowledge, such as Word-Net and Tagger. The most representative work in this discourse is Deerwester (1990) (FL=0.67) [14], which introduces a new approach of automatic indexing based on latent semantic analysis to overcome the deficiency of term-matching retrieval. Meanwhile, some researchers already have attempted the domain (topic modelling) and obtained remarkable results based on different methods, such as Latent Dirichlet Allocation.

3.5. Measure of Word Frequency

The fifth research stream explains 7.1% of the total variance and focuses on the research of word frequency measure related to POS Tagging. Word frequency measure is the underlying foundation of language research. In this discourse, the research mainstream is about English word frequency norms, e.g., both the most influential [15] and representative [16] articles focus on dealing with the improvement of SubtlexUS Corpus. Meanwhile, word frequency measure for other languages is also a concern by researchers around the world, for example, Keuleers (2010) compiled a database of Dutch word and character frequencies based on a corpus of film and television subtitles [17]. Publications in this discourse commonly share the same research path: Develop/improve an updated lexicon database based on previous word norms.

3.6. Methods of POS Tagging

The sixth research stream explains 6.4% of the total variance. This discourse mainly deals with the underlying methodology of POS Tagging, including the theory foundation, algorithm and compiling of POS. Santorini (1990) published a technical report to address notating problems in POS tagging, which not only presents a POS list but also provides detailed annotations and presentations for the lexicon [18]. The most influential article (Toutanova et al., 2003) introduces a new POS tagger to efficiently improve the accuracy and reduce the error of automatically learned tagging output [19]. Brill’s (1995) work [20], the most representative article, presents a new rule-based method for automated learning of language knowledge and verifies its validity in a POS tagging case.

4. Research system

4.1. Temporal evolution

Figure 1 shows the major changes of research involved in POS Tagging since 2008s, which outlines the discourse development. More than 50% of early research (Before 2011) focused on Methods of POS Tagging, represented by Toutanova’s (2003) new POS Tagging with a Cyclic Dependency Network [19] and Petrov’s (2006) automatic approach [21] to tree annotation. Meanwhile, current data on citations showed that topic modelling also acquired some attention since 2008. Although the discourse of topic modelling doesn't account for a large amount (less than 10%) in the entire field, the interest in it is continuing. Whereas the relative citation share of Methods of POS Tagging research has declined in volatility in recent years.

The intellectual structure has been broadened constantly since 2010. State-of-the-art models and Sentiment classification are gradually beginning to be concerned by researchers. With the emergence of novel models and architectures, the former began to pick up speed from 2016 and became the most dominant discourse (more than 60%) in the field nowadays. While the latter gradually shifted to another advanced branch called Aspect-based sentiment, and was rapidly being replaced by this novel branch since 2018. The research on Word Frequency, emerged rapidly in 2011 but a visible afterglow lingered briefly in 2012 and 2013 and decayed rapidly two years later.

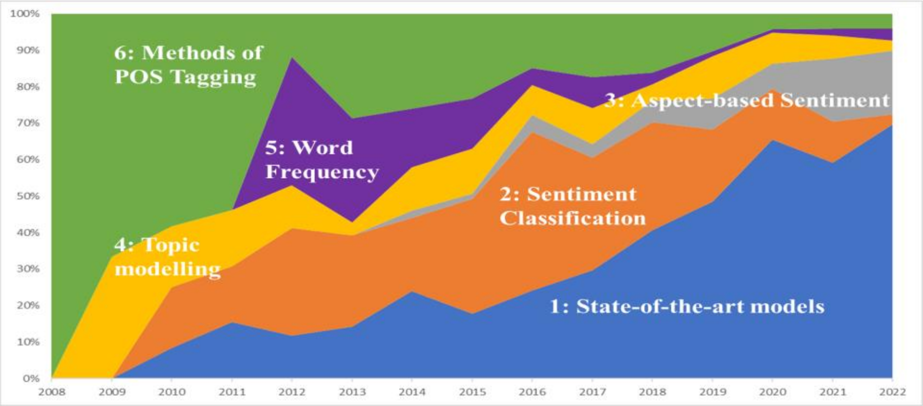

Figure 1: Prevalence of research streams over time.

4.2. Research network

In this section, we use social network analysis to uncover concealed structures between POS research streams. Figure 2 shows the similarities and differences in content between articles and research streams. Co-citations are linked by lines, and similar publications/streams are drawn closer together, while publications with higher cited frequency are represented by bigger nodes.

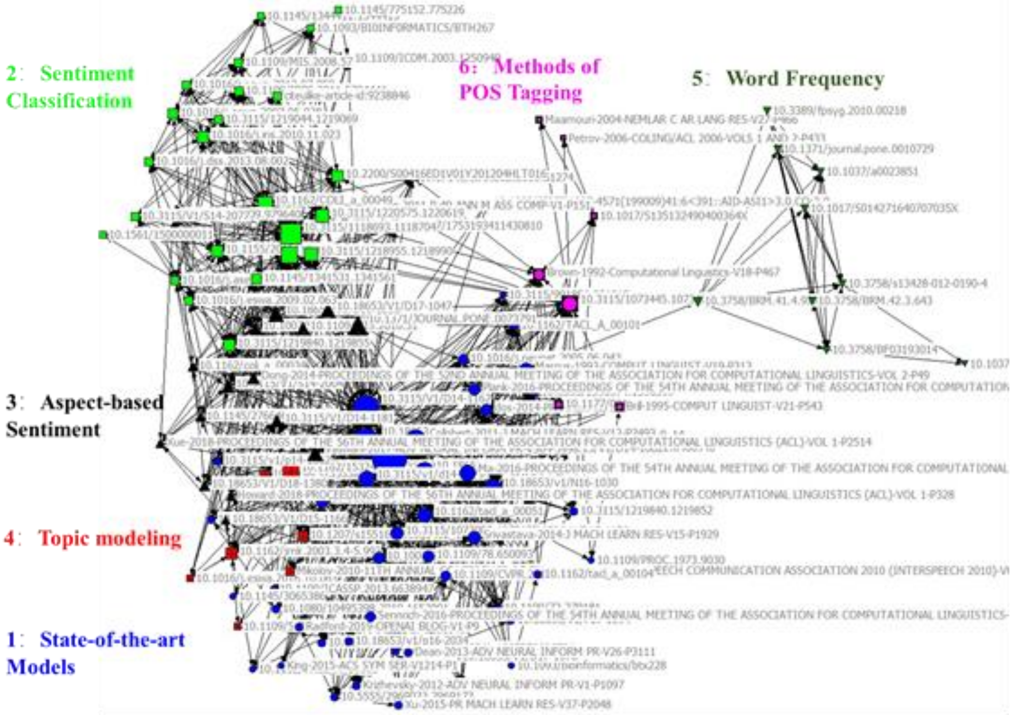
2: Sentiment Classification
6: Methods of POS Tagging
5: Word Frequency
3: Aspect-based Sentiment
4: Topic modeling
1: State-of-the-art Models

Figure 2: Intellectual structures of research streams.

When we examine the exchange of ideas between various streams in the research system, we can derive insights from the geographic proximity in Figure 2. Stream 5 (Word frequency measuring) operates as a distinct discourse, separate from the core research system. Stream 6 (POS Tagging application), is located at the center of the research system, maintaining connections with all other discourses while remaining relatively independent. Specifically, some of the studies in Stream 1 (State-of-the-art methods) and Stream 6 (POS Tagging application) have stronger idea exchanges, such as Marcus's (1993) [22] and Toutanova et al. (2003) studies [19].

Stream 1 (State-of-the-art methods), being the largest stream in the research system, has intensive collaboration with half of the research streams, particularly with Stream 3 (Aspect-based sentiment analysis) and Stream 4 (Automatic indexing and modelling). For instance, some of the studies in both Stream 1 (e.g., Devlin et al., 2019) [23] and Stream 3 (e.g., Chen et al., 2017) [24] share and exchange the novel idea of attention algorithm, while knowledge for learning vector space representations of words is shared and exchanged between Stream 1 (e.g., Devlin et al., 2019) [23] and Stream 4 (e.g., Bengio et al., 2000) [25]. Stream 2 (Sentiment classification) and Stream 3 (Aspect-based sentiment analysis) are heavily intertwined (e.g., Wang et al., 2016) [25]. Interestingly, Stream 3 (Aspect-based sentiment analysis) partly plays a linking role between Stream 2 and Stream 4, bridging the gap between sentiment classification and automatic indexing and modelling through “aspect-based sentiment analysis,” which involves both entity and sentiment detecting.

5. Discussions and conclusions

This study employs a bibliometric approach to provide an accessible understanding of POS Tagging knowledge in 921 WoS journal articles as well as their 40,647 cited references published from 2008 to 2022. Factor analysis on the co-citation matrix of 221 top-cited references reveal that POS Tagging shaped six crucial research streams. Earlier research mainly focused on the underlying approaches of POS Tagging and its basic applications such as word frequency calculation and topic modelling, but research interests have transferred to advanced applications of POS Tagging such as sentiment classification and novel NLP models in the last ten years. Our findings indicate that the major application based on POS tagging is sentiment analysis (including Aspect-based Sentiment analysis), while the State-of-the-art models are a focus in current academic research and may continue to play the lead in the future.

With the drastic development of state-of-the-art NLP technique, the efficiency of POS tagging has been significantly enhanced. The problem of manual POS tagging is laborious, while automatic labeling is knocked up against formidable difficulties, such as ambiguous words and unknown words. Researchers have continuously attempted to develop novel labeling models to improve the accuracy of POS tagging in recent years. Wherein, the labeling accuracy of most taggers has exceeded 96% [26]. However, 96% of the words or tokens may be not perfect enough. If a sentence consists of 20–30 words on average, a 96% accuracy implies that one term will be erroneously tagged per sentence. This error may affect other constituents and the following processing as a parser. We therefore continue to expect more state-of-the-art effort on automatic POS tagging.

References

[1] Abney, S. (1997). Part-of-Speech Tagging and Partial Parsing. In Corpus-based methods in language and speech processing (pp. 118-136). Dordrecht: Springer Netherlands.

[2] Martinez, A. R. (2012). Part-of-speech tagging. Wiley Interdisciplinary Reviews: Computational Statistics, 4(1), 107–113.

[3] Broadus, R. N. (1987). Toward a definition of “bibliometrics.” Scientometrics, 12(5–6), 373–379.

[4] Nerur, S. P., Rasheed, A. A., & Natarajan, V. (2008). The intellectual structure of the strategic management field: An author co-citation analysis. Strategic Management Journal, 29(3).

[5] Teichert, T., & Shehu, E. (2010). Investigating Research Streams of Conjoint Analysis: A Bibliometric Study. Business Research, 3(1), 49–68.

[6] Hota, P. K., Subramanian, B., & Narayanamurthy, G. (2020). Mapping the Intellectual Structure of Social Entrepreneurship Research: A Citation/Co-citation Analysis. Journal of Business Ethics, 166(1), 89–114.

[7] Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. Neural Computation, 9(8), 1735–1780.

[8] Ling, W., Lu ís, T., Marujo, L., Astudillo, R. F., Amir, S., Dyer, C., Black, A. W., & Trancoso, I. (2015). Finding function in form: Compositional character models for open vocabulary word representation. Conference Proceedings - EMNLP 2015: Conference on Empirical Methods in Natural Language Processing, September, 1520–1530.

[9] Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine learning techniques. Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - EMNLP '02, 10, 79–86.

[10] Subrahmanian, V. S., & Reforgiato, D. (2008). AVA: Adjective-Verb-Adverb Combinations for Sentiment Analysis. IEEE Intelligent Systems, 23(4), 43–50.

[11] Wang, Y., Huang, M., Zhao, L., & Zhu, X. (2016). Attention-based LSTM for aspect-level sentiment classification. EMNLP 2016 - Conference on Empirical Methods in Natural Language Processing, Proceedings, 606–615.

[12] Fan, F., Feng, Y., & Zhao, D. (2018). Multi-grained attention network for aspect-level sentiment classification. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018, 3433–3442.

[13] Bengio, Y., Ducharme, R., & Vincent, P. (2001). A neural probabilistic language model (short version). Advances in Neural Information Processing Systems.

[14] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. Journal of the American Society for Information Science, 41(6), 391–407.

[15] Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. Behavior Research Methods, 41(4), 977–990.

[16] Brysbaert, M., New, B., & Keuleers, E. (2012). Adding part-of-speech information to the SUBTLEX-US word frequencies. Behavior Research Methods, 44(4), 991–997.

[17] Keuleers, E., Brysbaert, M., & New, B. (2010). SUBTLEX-NL: A new measure for Dutch word frequency based on film subtitles. Behavior Research Methods, 42(3), 643–650.

[18] Santorini, B. (1990). Part-of-Speech Tagging Guidelines for the Penn Treebank Project (3rd Revision). University of Pennsylvania 3rd Revision 2nd Printing.

[19] Toutanova, K., Klein, D., Manning, C. D., & Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, HLT-NAACL 2003, June, 173–180.

[20] 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.

[21] Petrov, S., Barrett, L., Thibaux, R., & Klein, D. (2006). Learning accurate, compact, and interpretable tree annotation. COLING/ACL 2006 - 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, 1(July), 433–440.

[22] Marcus, M., Santorini, B., Ann Marcinkiewicz, M., & Large, B. (1993). Building a Large Annotated Corpus of English: The Penn Treebank Building a Large Annotated Corpus of English: The Penn Treebank Recommended Citation Recommended Citation. Computational Linguistics, 19(October), 313.

[23] David M. Blei, Andrew Y. Ng, M. I. J. (2003). Latent Dirichlet Allocation: Extracting Topics from Software Engineering Data. J Mach Learn Res, 3, 993–1022.

[24] Chen, T., Xu, R., He, Y., & Wang, X. (2017). Improving sentiment analysis via sentence type classification using BiLSTM-CRF and CNN. Expert Systems with Applications, 72, 221–230.

[25] Bengio, Y., Ducharme, R., & Vincent, P. (2001). A neural probabilistic language model (short version). Advances in Neural Information Processing Systems, 13

[26] Wang, Y., Huang, M., Zhao, L., & Zhu, X. (2016). Attention-based LSTM for aspect-level sentiment classification. EMNLP 2016 - Conference on Empirical Methods in Natural Language Processing, Proceedings, 606–615.