Proceedings of the 2nd International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/109/20241241

An Analysis of Current Situation and Prospect of Natural Language Processing Application in Discourse Analysis

Tianshuo Liu

School of Foreign Languages, Beihang University, Beijing, 100191, China

lts151932@icloud.com

Abstract. Natural Language Processing is a subfield of Artificial Intelligence and an interdisciplinary field of Linguistics and Computer Science. Previous research mainly focuses on the comparison and practical application of natural language processing technology, but the amount of research on its applications in Linguistics is limited. To fill this gap, this article focuses on the current application of technologies in discourse analysis, which is one of the most popular fields in Linguistics. The researcher will initially summarize the existing techniques and tools of Natural Language Processing used for pre-processing, such as Tokenization, Stemming, and Named Entity Recognition. Open-source tools such as NLTK and Spacy for pre-processing are also mentioned in the article. Subsequently, the researcher will introduce commonly used technologies like sentiment analysis and semantic analysis. This research concludes that discourse analysis can benefit many industries such as Medicine, Economics, Agriculture and Politics. However, this paper also analyzes the drawbacks and challenges that NLP technologies for discourse analysis face. Barriers such as Word Sense Disambiguation (WSD) and the context dependency of human language made it difficult for NLP technologies to be applied in Pragmatic analysis.

Keywords: Natural Language Processing, Discourse Analysis, Sentiment Analysis, pre-processing, Word embedding.

1. Introduction

Recently, there has been remarkable growth in research and development in Artificial Intelligence and its related areas. Natural language processing is another important subfield of artificial intelligence which endeavors to provide the computer with the ability to understand human spoken or written language. NLP, an interdisciplinary study of linguistics and computer science, involves the challenging task of enabling computer systems to understand human languages and interact with human beings with natural languages. Based on the presentation form of language, natural language processing technology can be applied to speech processing and text processing. Voice assistants like Siri and Xiao Ai have made interactions between human beings and robots a reality. These technologies not only provide convenience in daily life but also contribute to linguistic research like discourse analysis, which can explore real-world patterns and predict future trends infields like agriculture, healthcare, and politics. The development of advanced NLP technologies has significantly promoted interdisciplinary research between computer science and linguistics, providing opportunities for linguists to conduct quantitative research and make their analyses more convincing and well-founded. However, the applications of NLP

DOI: 10.54254/2755-2721/109/20241241

in discourse analysis seem to be limited to sentiment analysis in recent years. To probe into the status quo of the question, this paper will contain two tasks. Firstly, this paper will review the basic techniques of NLP based on previous study. Then, the applications of NLP techniques in discourse analysis, a key domain of linguistics, will be systematically introduced. Based on the preceding analysis, this paper aims to summarize the current state and identify the drawbacks of NLP applications in discourse analysis, offering suggestions for future development. This paper can provide linguistic researchers with an overview of NLP tools applicable in their research and offer future research directions for NLP techniques developers.

2. Natural Language Processing Preprocessing Techniques and Tools

2.1. Morphological processing

After a collection of text data is gathered, it cannot be immediately analyzed by large language models due to the presence of significant noise. Thus, the data must be cleaned in a stage called pre-processing. In this stage, Tokenization and Stemming are frequently used. Tokenization, also known as word segmentation, involves identifying words and dividing them into string sequences (commonly called tokens or words) based on specific criteria to facilitate further processing and analysis. There are many open-source tools such as NLTK and SpaCy. However, words with affixes can confuse the model, necessitating Stemming—a process that removes these affixes to standardize the words [1].

After the raw data is processed, researchers can analyze it from the perspective of vocabulary, a fundamental approach that offers insights into the types of words used and how they align with expected language norms[1]. There are also some tools for counting word frequency, showing a certain word's familiarity to speakers of the language[2], while Named Entity Recognition (NER) aims to identify and classify named entities in the text, such as people's names, place names, organization names, dates, times, and currencies[3]. The goal of NER is to mark out the entities in the text and classify them into predefined categories. This is critical for many applications and tasks such as information extraction, question-answering systems, machine translation, etc. Besides, parts of speech also provide information about the relative use of grammatical word classes. Automatic part of speech 'taggers' are based on the principle of 'sequence labeling problems'.[4] For example, the Natural Language Processing Toolkit (NLTK) uses machine learning algorithms, such as the Perceptron tagger, to learn context-specific tagging from a large corpus and then apply this knowledge to tag words in a sample [5].

In discourse analysis, researchers often use a related corpus, or one they have constructed themselves, containing billions of words. Thus, those simple and intuitive NLP techniques and tools make lexical analysis more accurate and efficient.

2.2. Word Meaning Processing

It is not enough to merely focus on the frequency and part of speech of words in a text dataset. To perform in-depth discourse analysis, it is essential to consider the meaning of words. However, it is extremely challenging to enable computers and machines to fully understand human language. Thus, the meaning of words should be transformed into vectors which can be processed by computers. Word embedding is a real-valued vector representation of words by embedding both semantic and syntactic meanings obtained from an unlabeled large corpus[6]. With advancements in NLP, various techniques for word embedding have emerged. Wang and Chen categorized the existing word embedding methods based on their underlying techniques. They divided those methods into seven categories, Neural Network Language Model, Continuous-Bag-of-Words and skip-gram, Co-occurrence matrix, FastText, N-gram model, Dictionary model and Deep contextualized model. There are also many popular baseline models introduced such as word2vec and GloVe. However, no word embedding model consistently performs well across all tasks. Thus, researchers should select models based on the characteristics of their datasets and the requirements of specific tasks[6].

DOI: 10.54254/2755-2721/109/20241241

3. Applications of NLP Techniques in Discourse Analysis

3.1. Semantic Analysis

3.1.1. Introduction of Semantic Analysis. Semantic analysis is a specialized technique within the broader field of research that merges NLP with Discourse Analysis. This area intersects with artificial intelligence, opinion mining, text clustering, and classification. When linguists conduct discourse analysis, the meaning of context plays a vital role in conveying ideas and thought of people. Thus, Latent Semantic Analysis (LSA) technologies can be applied, which leverages word co-occurrence information from a large unlabeled corpus of text. Unlike methods that rely on explicit human-organized knowledge, LSA "learns" its representation by applying Singular Value Decomposition (SVD) to the word-by-document co-occurrence matrix[7]. In his research, Gabrilovich proposed a new technique called Explicit Semantic Analysis (ESA), which uses knowledge concepts explicitly defined and manipulated by humans[7].

3.1.2. Applications of Semantics Analysis. Semantic analysis is widely applied in different fields, such as Agriculture, Medicine, and Economics. It can process opinions and ideas posted online, contributing valuable insights to specific fields.

In the year 2022, Mehak Rehman and his partner conducted research providing a descriptive analysis of collected agrarian experts' opinions to increase crop yield by NLP semantic analysis techniques. They used semantics analysis on agriculture datasets to explore similarities, sentiments, emotions, feelings, and thoughts regarding crop productivity [8]. In addition to the preprocessing techniques mentioned in the previous section, they also used Naïve Bays (NB) and K Nearest Neighbors (KNN) models for classification. Their study not only proved that the machine learning algorithm Naïve Byes performs better on text datasets but also showed that Crop, Certified Seed, and Post-Harvest Loss significantly contribute to agricultural productivity.

Semantic analysis can be applied in medical research, as understanding semantic memory and language can reflect a person's mental status. By analyzing patients' daily discourse, doctors can detect early signs of cognitive or neurological disorders. In 2017, Gliner scored semantic fluency in patients with autism spectrum disorders. In 2019, Hoffman applied LSA to illuminate the neural systems supporting coherent speech in healthy adults. In the Alzheimer's disease (AD) field, Dunn, Almeida, Barclay, Waterreus, and Flicker found that using LSA word embeddings to compute the similarity between patients' attempts at a story recall and the original passage outperformed traditional hand scoring.[9]

3.1.3. Research Gaps and Challenges in Semantics Analysis. Although semantic analysis techniques like LSA and ESA have made significant contributions to discourse analysis and various practical fields, some challenges remain to be addressed. One major issue is Word Sense Disambiguation (WSD), which continues to be a difficult and critical problem in natural language processing. Unique sentences in a discourse such as idioms, sarcasm, or cultural references cannot be accurately detected and analysed. Thus, if the problem of WSD cannot be solved, the gap between semantics and pragmatics, or integrating real-world knowledge and context cannot be filled. Current research in pragmatics largely relies on theoretical case analysis. Therefore, analyzing the meaning of words or sentences according to different contexts is a critical challenge for developing truly comprehensive semantic analysis.

3.2. Sentiment Analysis

3.2.1. Introduction of Semantic Analysis. Sentiment analysis is a process of automatically identifying whether text data expresses positive, negative or neutral opinions about a topic or entity. Sentiment analysis, as a part of discourse analysis, has played an important role in the fields of politics, economics, business, and media. Extensive research has been conducted on sentiment analysis for movie reviews

DOI: 10.54254/2755-2721/109/20241241

and news articles, and many sentiment analyzers are available as open source[10]. Popular machine learning algorithms like Naïve Bayes and Decision are used for sentiment analysis.

3.2.2. Applications of Sentiment Analysis. Due to the availability of text datasets, many researchers conduct sentiment analysis based on social media like Twitter to explore public ideas and thoughts. For instance, in 2012 Wang and his partner built a baseline sentiment model using Amazon Mechanical Turk, which aimed to conduct a real-time analysis of public sentiment toward presidential candidates in the 2012 U.S. elections. Compared to traditional discourse analysis, which generally takes weeks, this system can instantly and continuously output the results according to dynamics of public events and election process. And this model cannot only react to public events like elections, but also can be applied to other fields like movie reviews and nominations for film and television awards[11]. Besides, Pagolu and his team analyzed the correlation between a company's stock market movements and sentiments expressed in tweets. They used two different techniques, Word2vec and N-gram to analyze the sentiment of people's comments and tweets posted on Twitter. In the research, they found that positive emotions or sentiments of the public on Twitter about a company would be reflected in its stock price.

3.2.3. Research Gaps and Challenges in Sentiment Analysis. There are numerous models pre-trained for sentiment analysis, but the output results are discrete types, which means that those models can only predict if a discourse is positive, neutral or negative rather than truly analyzing the specific emotions of the discourse. Unlike traditional emotion analysis, linguists can accurately identify people's emotions based on words used in specific contexts, such as satisfaction, jealousy, and helplessness. However, sentiment analysis based on a large language model can only give prediction results according to the annotation of its training data set. If the data labels in the training set are only positive, neutral, and negative, the large model is unable to further predict the specific emotion of the discourse. To achieve more in-depth sentiment analysis, overcoming the limitations of semantic analysis is essential.

4. Conclusion

The underlying frameworks and tools for natural language processing are well established. However, processes like tokenization, stemming, and word embedding are the foundation of natural language analysis, not the ultimate goal. Nowadays, the application of machine learning and deep learning has gradually integrated natural language processing technology into discourse analysis, such as sentiment analysis and semantic analysis. The development of Natural Language Processing has restructured methods of discourse analysis by making them more efficient and data-dependent. However, there are still limitations that need to be overcome. In the field of discourse analysis, researchers have found that NLP technology analysis lacks the embedding of linguistic theory. Most of the research focuses on collecting large amounts of data, using data sets to train models, and analyzing prediction results, without in-depth discussion of the linguistic meaning of discourse. The analysis of discourse comes from linguistics. Relying solely on algorithms leads to the homogenization of analysis methods, making them unsuitable for studying pragmatics. Thus, the context dependency of human language and the phenomenon of polysemy are the most significant challenges. In the future, researchers should prioritize solving the problems of WSD and strive to make pragmatic analysis with NLP techniques a reality, if they want to integrate pragmatics into semantic models and move towards a more comprehensive and contextually aware NLP era.

References

[1] Nagarhalli, T. P., Vaze, V., & Rana, N. K. (2021, February). Impact of machine learning in natural language processing: A review. In 2021 third international conference on intelligent communication technologies and virtual mobile networks (ICICV) (pp.1529-1534). IEEE.

[2] Balota, D. A., & Chumbley, J. I. (1984). Are lexical decisions a good measure of lexical access? The role of word frequency in the neglected decision stage. Journal of Experimental Psychology, 10(3), 340e357.

DOI: 10.54254/2755-2721/109/20241241

[3] Agerri, R., Bermudez, J., & Rigau, G. (2014, May). IXA pipeline: Efficient and Ready to Use Multilingual NLP tools. In LREC (Vol. 2014, pp. 3823-3828).
[4] Marcus, M. P., Marcinkiewicz, M. A., & Santorini, B. (1993). Building a large annotated corpus of English: The Penn treebank. Computational Linguistics, 19(2), 313e330.
[5] Bird, S., Klein, E., & Loper, E. (2009). Natural Language Processing with Python: Analyzing text with the Natural Language toolkit. O'Reilly Media, Inc.. http://nltk.org/book
[6] Wang B, Wang A, Chen F, Wang Y, Kuo C-CJ. Evaluating word embedding models: methods and experimental results. APSIPA Transactions on Signal and Information Processing. 2019;8:e19. doi:10.1017/ATSIP.2019.12
[7] Gabrilovich, E., & Markovitch, S. (2009). Wikipedia-based semantic interpretation for natural language processing. Journal of Artificial Intelligence Research, 34, 443-498.
[8] Rehman, M., Razzaq, A., Baig, I. A., Jabeen, J., Tahir, M. H. N., Ahmed, U. I., ... & Abbas, T. (2022). Semantics analysis of agricultural experts' opinions for crop productivity through machine learning. Applied Artificial Intelligence, 36(1), 2012055.
[9] Clarke, N., Foltz, P., & Garrard, P. (2020). How to do things with (thousands of) words: Computational approaches to discourse analysis in Alzheimer's disease. Cortex, 129, 446-463.
[10] Pagolu, V. S., Reddy, K. N., Panda, G., & Majhi, B. (2016, October). Sentiment analysis of Twitter data for predicting stock market movements. In 2016 international conference on signal processing, communication, power and embedded system (SCOPES) (pp. 1345-1350). IEEE.
[11] Wang, H., Can, D., Kazemzadeh, A., Bar, F., & Narayanan, S. (2012, July). A system for real-time Twitter sentiment analysis of 2012 us presidential election cycle. In Proceedings of the ACL 2012 system demonstrations (pp. 115-120).