Extraction and Quantitative Analysis of Core Vocabulary in Maritime English for Vocational Education

Kang Liu

Emmail:lk@zjvtit.edu.cn

Navigation College, Zhejiang Institute of Communications. Hangzhou, Zhejiang 311112

Abstract: For a long time, the focus of maritime English reform has been on undergraduate students, with limited research on maritime English reform for vocational college students. However, the vocabulary mastery level of vocational college students directly affects their learning quality. This study uses vocational maritime English and related teaching materials as the basis, and constructs a corpus of vocational maritime English with 8,192,189 words using the NLTK library. With the Word2vec model, unstructured text is tokenized and embedded in the word vector space, converting the text into numerical values with semantic information. Through clustering, statistical analysis, and feature extraction, the study analyzes the high-frequency vocabulary, part-of-speech distribution, sentence length distribution, and sentiment tendencies in vocational maritime English. The results show that the constructed corpus and trained word vectors can effectively record semantic information. The findings of this study can provide teaching references for vocational maritime English, improve teaching effectiveness, and offer insights for relevant educational reforms.

Keywords: Maritime English, core vocabulary, teaching reform

1 Introduction

For a long time, teaching reforms have been carried out in Maritime English. Scholar Han Guochong[1] proposed in his research and practice on the reform of English teaching in maritime majors in universities that "effective reforms are needed in the English teaching of maritime majors in China to meet the requirements of the modern maritime industry for high-quality crew members with proficient English skills.". Scholar Wang Fang[2] conducted research on maritime English at the undergraduate stage and believed that "the vocabulary density and average word length of maritime English in universities are smaller than those of ordinary English. The high-frequency words in maritime English cover more maritime-related vocabulary or closely related vocabulary", and proposed that maritime English should adopt "content-based teaching.".

The process of teaching reform in Maritime English had made some progress, especially in recent years, with the integration of the Internet and the introduction of advanced teaching methods such as MOOCs and micro-lessons, which had improved teaching effectiveness to some extent[3]. However, there was relatively little research on the teaching of Maritime English in vocational colleges. Vocational college students could not be taught English in the same way as undergraduate students, but vocational college students were the main source of inter-

ICEKIM 2023, May 26-28, Nanjing, People's Republic of China

DOI 10.4108/eai.26-5-2023.2337282

national seafarers in China. In terms of vocabulary teaching, the characteristics of vocational college students, such as their limited vocabulary, weak English knowledge framework, and strong professional requirements for Maritime English, had been taken into consideration. Therefore, emphasis had been given to key vocabulary, sentence structures, and content to strengthen their learning. In this paper, Maritime English core vocabulary was extracted, classified, quantified, and analyzed using Python programming language based on Maritime English textbooks and related materials, in an attempt to form structured knowledge content, which was expected to have a positive and effective impact on actual teaching.

2 Model Architectures

This paper uses the Python programming language to extract, classify, quantify, and analyze core maritime English vocabulary by creating a maritime English corpus. For vocabulary analysis, the word2vec model[4]-[6] converts words into word vectors for easy processing by computers, resulting in optimized text analysis.

The natural language processing usually represents words as single, discrete numbers and often uses One-hot Representation to represent words as high-dimensional vectors consisting of 0 and 1[7]-[10]. This method cannot fully express the semantic information of the words and presents difficulties in computation due to high-dimensional calculations. Word2vec is a simplified neural network model based on NNLM, consisting of only three layers: input, hidden, and output, which facilitates text processing for small corpora. There are two types of model frameworks based on different input-output:

2.1 Continuous Bag-of-Words Word2Vec(CBOW)

The CBOW model is an algorithm for the derivation of word vectors, which is targeted at the word vectorization model. The model understands the $context(w)$ of the text to deduce the word vector of each vocabulary w_t in the text. The CBOW algorithm is widely used in predicting the next vocabulary, text classification, and text correlation. Its schematic diagram is as Fig. 1:

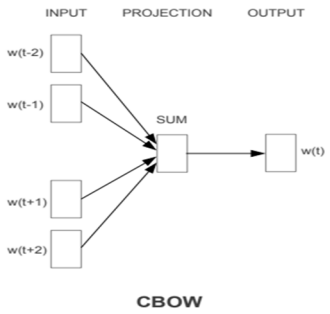

Fig. 1. CBOW model

The objective function for CBOW is:

$$J_\theta = \frac{1}{T}\sum_{t=1}^{T} logp(w_t \mid w_{t-n}, \ldots, w_{t-1}, w_{t+1}, \ldots, w_{t+n})$$

2.2 Skip-gram

The Skip-gram mode is an algorithm that infers the overall structure of the text. This mode deduces the vector of each vocabulary $context(w)$ in the text by identifying the key vocabulary w_t in the text. The Skip-gram mode is highly favored for its ability to handle complex text structures and is widely used in text analysis, text classification, and text formal analysis. Its schematic diagram is as Fig. 2:

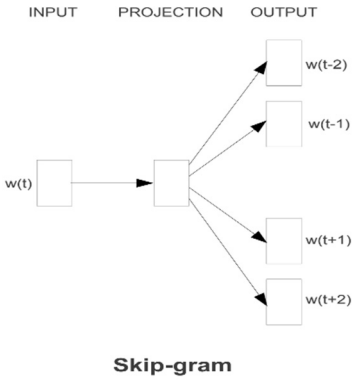

Fig. 2. Skip-gram mode

The objective function for CBOW is:

$$J_\theta = \frac{1}{T}\sum_{t=1}^{T} \sum_{-n \le j \le n, \neq 0} \log\, p(w_{j+1} \mid w_t)$$

In this paper, we will use the word2vec model to extract the core vocabulary of vocational maritime English and quantitatively analyze the extracted core vocabulary. From the results, the skip-gram algorithm obtains more accurate results than CBOW. Therefore, this paper mainly presents the data results generated by the skip-gram mode.

3 Self-construction and processing of corpus

The key to this study was to establish a corpus that met the research objectives using programming tools. The corpus should reflect the research topic as broadly as possible while covering enough data to ensure the accuracy of the research results. After building the corpus, a series of processing was conducted to provide a foundation for information extraction and quantitative analysis.The constructions is as Fig. 3.

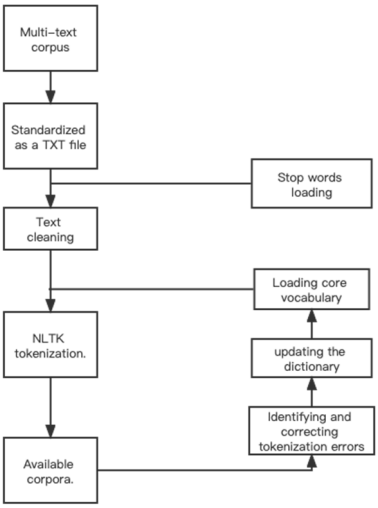

Fig. 3. The Construction Diagram of Corpus

3.1 Data Collection and Cleaning

To build a Maritime English corpus, we collected a large amount of English texts from explicit Maritime English textbooks, related materials, English academic papers, reports, news reports, blogs, social media articles, and other Maritime English text resources, and conducted preprocessing and cleaning. We used the NLTK library in Python for text preprocessing and tokenization. Preprocessing included removing punctuation, stop words, converting English letters to lowercase, removing stop words (such as "the", "a", etc.), and words with a length less than three. In consideration of the professional characteristics, we used an expert mode to annotate some special terms. Based on this, we conducted tokenization to obtain a cleaner dataset.

3.2 Building a Dictionary

To better reflect the characteristics of marine technical English, we used the NLTK library to conduct preliminary frequency statistics on the processed corpus, and selected the top 15,000 words with the highest frequency as the dictionary. At the same time, we filtered out low-frequency words and stop words again, only retaining meaningful high-frequency words. Finally, we saved the resulting corpus for later revisions and continued research.

3.3 Establishing a Co-occurrence Matrix

Based on the constructed dictionary, we established a co-occurrence matrix, which was a tool for analyzing the correlation between words in the text. It constructed a matrix by counting the number of times words appeared together in the article. Each element in the matrix represented the number of times two words co-occurred. This matrix could help us with multiple natural language processing tasks, such as text clustering, sentiment analysis, and information retrieval. In this study, we denoised the co-occurrence matrix, performed text clustering on the Maritime English corpus, and conducted sentiment analysis, striving to obtain useful text information.

3.4 Using Gensim Library for Word2Vec Computation

Using the Word2Vec model in the Gensim library and the co-occurrence matrix calculated in the previous step, we analyzed and extracted the core vocabulary of Maritime English. By calculating the Word2Vec vector corresponding to each word, we could find the top 100 words that were most similar to that word. These words could be considered the core vocabulary of Maritime English, which was important for Maritime English teaching.

Through our self-built Maritime English corpus, we could extract and analyze the core vocabulary of Maritime English, obtain relevant statistical information and knowledge content. This would help us better understand and apply Maritime English, providing powerful support for work and research in the marine field.

4 Extraction and Analysis of Core words

Building on the above, this study used a self-built Maritime English corpus to extract and quantitatively analyze the core vocabulary of Maritime English, as shown below:

4.1 1Extraction of the Core Vocabulary of Maritime English

Through the self-built corpus and using the Word2Vec model in the Gensim library, we found the top 100 words that were most similar to each word in the corpus. These words could be considered the core vocabulary of Maritime English (as shown in Fig. 4). We arranged these core vocabulary words according to their frequency of occurrence, and obtained a list of the core vocabulary of Maritime English. These words covered various knowledge and concepts in the marine field, which was of great significance for understanding Maritime English and related work.

ship(1152)	area(181)	side(135)	speed(113)	rules(100)
cargo(570)	passage(179)	emergency(133)	mariners(112)	stability(100)
vessel(477)	position(170)	made(131)	anchor(112)	case(99)
may(318)	also(159)	oil(129)	given(111)	bridge(99)
ships(298)	master(156)	ensure(126)	provided(111)	space(98)
sea(296)	board(155)	control(125)	container(109)	safe(98)
system(263)	navigation(154)	company(125)	lifeboat(109)	conditions(98)
water(260)	two(152)	vessels(124)	data(106)	full(96)
used(241)	surface(151)	international(123)	general(106)	radar(96)
port(239)	officer(145)	means(123)	winds(106)	cargoes(96)
fire(239)	weather(144)	hours(121)	engine(106)	drill(96)
deck(236)	hold(143)	loading(119)	convention(105)	line(95)
chart(221)	part(139)	watch(118)	force(104)	notices(94)
must(196)	operation(139)	condition(118)	requirements(104)	range(94)
one(196)	charts(138)	code(117)	light(103)	necessary(94)
information(195)	following(138)	navigational(116)	discharge(103)	good(94)
safety(195)	security(138)	less(115)	fog(102)	tonnage(94)
crew(192)	air(137)	admiralty(114)	within(102)	test(94)
equipment(190)	wind(137)	number(114)	list(100)	carried(92)
time(183)	required(136)	state(114)	main(100)	special(92)

Fig. 4. core vocabulary of Maritime English

4.2 Statistical Analysis of Part-of-Speech of Maritime English

We conducted part-of-speech tagging and statistical analysis on all the words in the corpus (as shown in Fig 5). From the statistical results, nouns (NN, NNS), adjectives (JJ), and gerunds (VBG) accounted for 42%, 33%, and 8% of the total quantity, respectively, and the three of them accounted for 90% of the total. It can be seen that nouns played a dominant role in Maritime English, and the content of Maritime English texts was mainly explanatory and narrative, with few emotional adjectives and other words. This requires students to focus on understanding and learning nouns, which can improve learning efficiency.

At the same time, we noted that there were about 3% of proper nouns (NNP) in the corpus, most of which were specialized marine vocabulary with strong professional characteristics. Although the total quantity was not large, considering that students were first exposed to specialized vocabulary, they should also be given special attention and priority in learning.

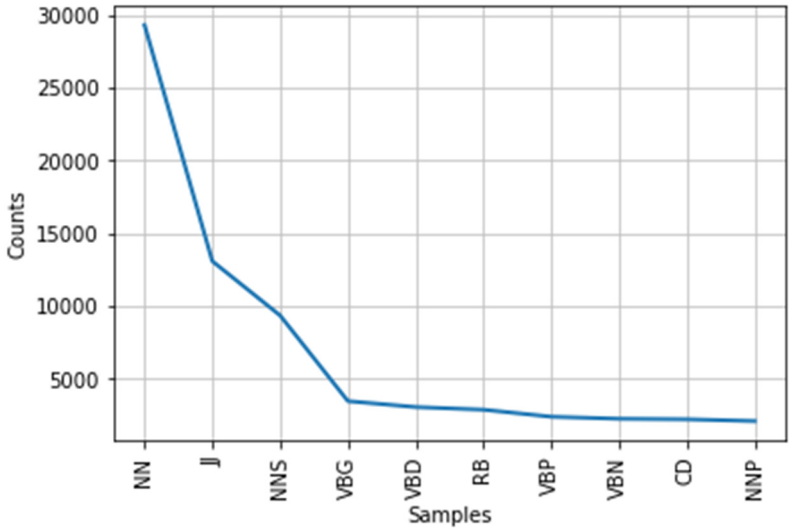

Fig. 5. part of speech statistics of Maritime English

4.3 Statistical Analysis of Text Length of Maritime English

On the basis of cleaning the text, we conducted statistical analysis of the content according to the text length (as shown in Fig 6). From the statistical results, short and medium sentences with no more than 20 words accounted for 93% of the Maritime English texts, while long sentences with more than 40 words accounted for only 3% of the total. Among the main content with no more than 20 words, sentences with no more than 10 words accounted for 65% of the total, occupying the main part. Therefore, we believe that Maritime English in vocational education mainly consists of short sentences with no more than 10 words, with few long sentences, which are mainly related to legal provisions or clauses related to maritime conventions according to the content.

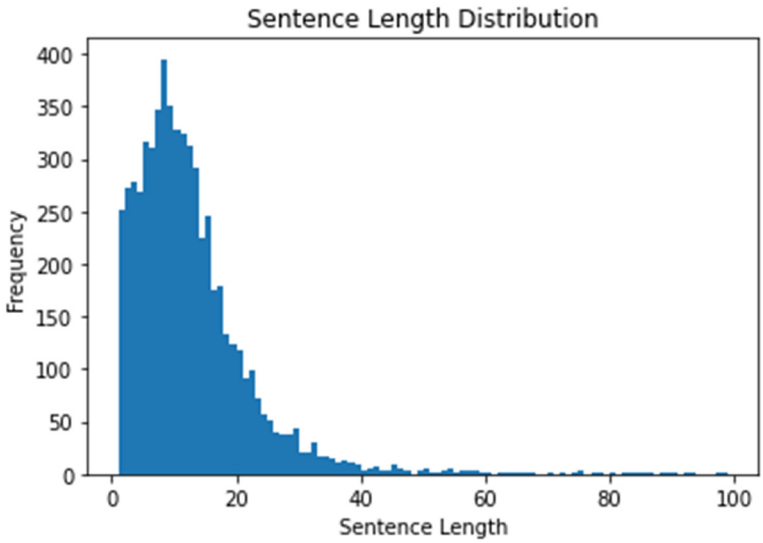

Fig. 6. length of Maritime English text

4.4 Preliminary Assessment of Sentiment Tendency in Maritime English

To further analyze the Maritime English corpus in vocational education, we attempted to obtain the sentiment tendency of the content. On the basis of cleaning the text, we used the Word2Vec word vector model to train the content and extracted its sentiment tendency through feature extraction (as shown in Fig 7). As can be seen from the figure, neutral sentiment content accounted for the majority, at 53%; negative sentiment content accounted for 41%, and positive sentiment content was relatively small. By referring to content materials, there is a large amount of safety-related content in Maritime English in vocational education, such as ship firefighting, personal survival, basic first aid, etc., which involves accident contents such as fire, sea disasters, collisions, and grounding. This is the main reason why negative content accounts for a relatively large proportion. At the same time, content related to ship instruments, navigation books, ship inspections, and other professional explanations are neutral, which, together with other specialized explanations, forms the main body of the neutral content.

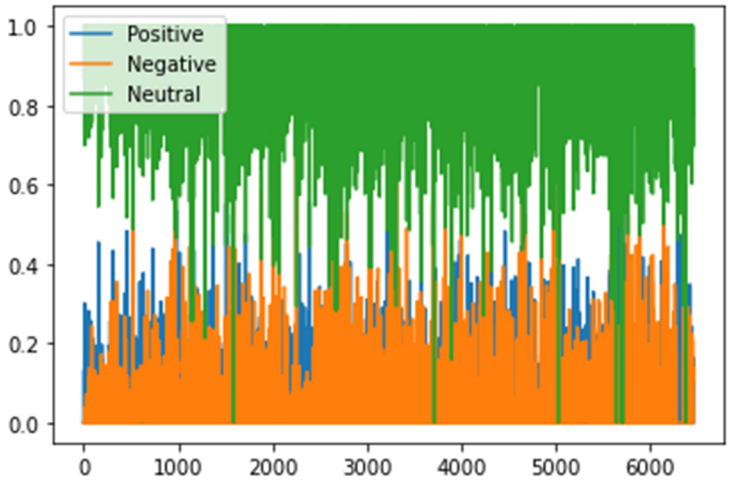

Fig. 7. navigational English emotional tendency

5 Discussion and summary

In this study, we built a self-built corpus and conducted cleaning and processing, extracted core vocabulary, analyzed part-of-speech distribution, presented text length distribution, and judged sentiment tendency, which showed the characteristics of Maritime English vocabulary in vocational education. From the results, we believe that:

1. Relevant vocabulary in the maritime field plays an important role and showcases the basic content of maritime-related work. In terms of vocabulary complexity, the complexity is not high, and most of them are three-syllable or fewer words. Therefore, the core vocabulary of Maritime English in vocational education mainly consists of basic vocabulary that conforms to the maritime field, which can be the focus of students' learning.
2. In terms of part-of-speech distribution, nouns occupy the main position. Compared with other English learning content, such as grammar and expression, Maritime English pays more attention to the embodiment of substantive content, which has an enlightening effect on teachers in teaching.
3. In terms of text length, most of the content involved in this study are short sentences with no more than 20 words, which is in line with the accuracy and conciseness of Maritime English. Considering the importance of legal conventions in the profession, although sentences with more than 40 words have little content, they should also be understood and mastered.
4. From the perspective of sentiment analysis, neutral sentiment predominates in Maritime English in vocational education, and negative sentiment is mainly due to the inclusion of more knowledge about ship safety. Therefore, the author recommends adding more positive sentiment content, such as professional honor and social responsibility, in future content supplements to encourage students to engage in the maritime industry and enhance their professional confidence.

6 Conclusion

This study aimed to extract core vocabulary and analyze the content of Maritime English in vocational education through the construction of a specialized corpus. The research found that Maritime English in vocational education was mainly related to maritime-related work, with relatively low vocabulary complexity, nouns as the main part-of-speech, moderate text length, and a focus on professional content. The distribution of part-of-speech was consistent with the results of Xu Xin's study, where nouns dominated [Xu Xin. Research on Maritime Vocabulary in English Maritime Films Based on Corpus Analysis[D]. Dalian Maritime University]. In the part of sentiment tendency, the author proposed their own views and suggestions. The research content of this article could serve as auxiliary material for students' self-study and provide reference for teaching Maritime English in vocational education. However, the method used in this study still had certain limitations, as it only analyzed individual discrete words and did not model and analyze phrases and sentences, which might have affected the accuracy of the results. In future research and application, we will further explore and improve this method to obtain more accurate, comprehensive, and efficient analysis results.

Acknowledgment: This research project was funded by the General Research Project of Zhejiang Provincial Department of Education, NO: 6230304Z91108.

Reference

[1] Han Guochong, Gu Xingming. Research and Practice on the Reform of English Teaching in Maritime College Majors. Journal of Jiamusi Vocational College, 2020, 36(12):3.

[2] Wang Fang. Quantitative Research on English Vocabulary in Maritime College and Teaching Implications. Journal of Maritime Education Research, 2014, 31(2):4.

[3] Liu Chang, You Haoqi, Long Dan, et al. Characteristics of Civil Aviation English Professional Vocabulary and the Construction of Civil Aviation Image from the Perspective of Self-built Corpus. Science and Technology Horizon, 2022(22):4.

[4] Jiang Yue. Extraction and Application of High-frequency Core Vocabulary in Mechanical Engineering English. Neijiang Science and Technology, 2021.

[5] Wang Fangshu. Analysis of Nautical Vocabulary in Nautical Novels Based on Corpus. [D]. Dalian Maritime University.

[6] Hao Xiaoyan, Shu Maoguo. Teaching Exploration of English Vocabulary in Plastic Surgery Major. Chinese Journal of Aesthetic Medicine, 2021, 30(1):4.

[7] Xu Xin. Study on Maritime Vocabulary in English Maritime Films Based on Corpus Analysis. [D]. Dalian Maritime University.

[8] Chen Lili. Cognitive Metaphor Teaching of Nautical English Vocabulary Based on Corpus. Journal of Maritime Education Research, 2018, 35(2):3.

[9] Dirgayasa I W . Maritime English Learning Materials Based on Standard Training Certification and Watchkeeping for Seafarers (STCW) Curriculum and Intercultural Competence (IC)[J]. World Journal of English Language, 2022, 12.

[10] Xu L . Assessing change in English second language writing performance. Khaled Barkaoui & Ali Hadidi New York/London: Routledge, 2020, Innovations in language learning and assessment at ETS, 218 pp. ISBN: 9780367551902 (pbk)[J]. International Journal of Applied Linguistics, 2021.