JDCS
디지털콘텐츠학회논문지
Journal of Digital Contents Society
Vol. 22, No. 1, pp. 71-75, Jan. 2021

Check for updates

영어 어휘 학습용 단어리스트 생성을 위한 번역 시스템 활용

우 진 하[1] · 최 희 열[2*]
[1]한동대학교 언어교육원 초빙교수
[2*]한동대학교 정보통신공학과 부교수

Utilizing Machine Translation Systems to Generate Word Lists for Learning Vocabulary in English

Jin-Ha Woo[1] · Heeyoul Choi[2*]
[1]Visiting Professor, Dept. of Language Education, Handong University, Pohang, Korea
[2*]Associate Professor, Dept. of Information and Communication Engineering, Handong University, Pohang, Korea

[요 약]

딥러닝 기반 번역 시스템의 역량과 접근성이 급격히 향상됨에 따라 상당수의 학생들이 다양한 언어 학습 목적으로 온라인 번역기를 활용하고 있다. 언어학습에서 온라인 번역기는 어휘 학습에 가장 많이 활용된다. 본 논문은 학습에 유용하고 적절한 어휘를 선택하기 위해 번역 시스템에서 입력되고 번역된 텍스트를 활용하여 개인화된 단어 목록을 생성할 것을 제안한다. 실험에서는 번역 시스템에 단어를 추출하는 모듈을 추가하고, 실제 사례를 보여주기 위해 '인공지능' 이라는 단어와 관련된 영어 50개 기사와 한국어 25개 기사의 번역된 영어 원문을 바탕으로 단어 목록을 작성했다. 제안된 방법을 통해, 널리 사용되는 다른 학술적 단어 목록과는 많이 겹치지 않고, 개별 번역기 사용자의 관심사와 직접적으로 일치하는 단어 리스트를 만들 수 있었다. 사용자에게 맞춤형 단어리스트를 제공함으로써 학생들은 특수한 영어 어휘 지식을 효과적으로 확장하기위해 유용한 단어들을 식별하고 전략적으로 공부할 수 있게 된다.

[Abstract]

Following a rapid increase in the capabilities and accessibility of neural machine translation (NMT) systems, a substantial number of students utilize online translators (OTs) for diverse language learning purposes. In language learning, OTs are frequently used to scaffold vocabulary learning. To select useful vocabulary words for learning, in this paper, we propose utilizing source and target text in NMT systems to generate personalized word lists. In the experiment, we implemented a word counting module in our NMT system and generated word lists based on source words from 50 English articles and target English words of 25 Korean articles with content related to Artificial Intelligence (AI) to show an example. Through the proposed method, we generated a highly specific word list that directly aligned with the interests of individual OT users and had low word overlap with other widely used academic word lists. With these personalized word lists, students can strategically identify and study useful words to effectively increase their specialized English vocabulary knowledge.

색인어 : 영어, 언어학습, 신경망 기반 번역 시스템, 어휘, 단어 리스트
Key word : English Language, Language Learning, Neural Machine Translation System, Vocabulary, Word Lists

http://dx.doi.org/10.9728/dcs.2021.22.1.71

Received 26 November 2020; **Revised** 05 January 2021
Accepted 05 January 2021

***Corresponding Author; Heeyoul Choi**

Tel: +82-54-260-1303
E-mail: hchoi@handong.edu

Ⅰ. Introduction

The capabilities and use of online neural machine translation (NMT) systems have been rapidly increasing as evidenced by Naver Papago [1] and Google Translate [2]. In 2019, Naver Papago reported over 10 million monthly active users [1]. In 2018, Google Translate reported translating “30 trillion sentences per year across 103 languages” [2].

An overwhelming majority of students utilize online translators (OTs) for language learning. Based on the results of large-scale questionnaires [3, 4, 5], 71% to 99% of students utilized OTs in their foreign language course. These students frequently used OTs to check words/phrases and perceived OTs as helpful for increasing target language vocabulary [3].

In this paper, we analyzed source and target text in a Korean-English NMT system to scaffold vocabulary learning. The source and target text was used to generate personalized English vocabulary lists based on the word frequency counts. These lists can help students to effectively study vocabulary words aligned with their academic goals. With these lists, students can strategically identify and study useful words needed to understand specialized texts and communicate technical ideas [6].

The rationale for our work is twofold. First, we understand that word selection is a crucial aspect of learning target language vocabulary. To select useful words, occurrence frequency must be considered as high-frequency words provide better returns for learning efforts [6]. Second, we noticed that users are unable to access word frequency statistics in widely used OTs. For Google Translate and Naver Papago, users can only select and save translated words/phrases in an online phrasebook.

To generate personalized word lists, we implemented a word counting module in our NMT system. First, we counted words in source and target sentences, respectively. After filtering out stop words [7] and high-frequency words [8], we generated two English vocabulary lists with the 100 most frequent words ranked from the highest to lowest frequency count. The first list was generated from 50 English articles with content related to the word ‘AI’ inputted into our NMT system. The second list was generated from 25 Korean articles with ‘AI’ content translated in our NMT system. When compared to other widely used academic word lists, there was low word overlap and these lists contained noticeably more area-specific words.

Ⅱ. Background: Translation Systems

In this section, we provide a brief overview of our Korean-English NMT system, which consists of preprocessing, translation, and postprocessing.

2-1 Translation Model

Since deep learning was applied to natural language processing, Recurrent Neural Networks (RNNs) were mainly used in the early days of NMT [9]. Since then, models using Convolution Neural Networks (CNNs) have been proposed [10], and recently, the Transformer using the attention mechanism has become the mainstream neural network-based translation model [11]. The Transformer model maintains the existing Encoder and Decoder structures as in RNN models, but uses only the attention mechanism to improve the shortcomings of the RNN-based model, resulting in faster learning time and improved performance. In this paper, the Transformer model is used as the basic translation model. For a comparison of the translation models, see [12].

2-2 Corpus and Preprocessing

In this paper, we used Korean-English corpus data provided at the ‘AI Hub’ by the National Information Society Agency, which includes the training data necessary for general translation work with 1.6 million pair-wise sentences. The data is split into two sets: 1,596,418 training data and 3,000 validation data.

To improve learning performance, preprocessing of data is performed according to the purpose. The first step is to split the sentence into several pieces of token units based on punctuation marks such as dots, commas, question marks, and quotation marks. The second step is to handle the OOV (Out of Vocabulary) problem. In natural language processing, the list of words is sometimes very long, which leads to the problem of increasing the size of the model. Even with a very long word list, it is difficult to learn unusual words. Moreover, it is not possible to translate OOV words that are not in the training corpus. To solve this problem, one of the most popular methods is to divide each word into several meaningful subwords, as in Byte Pair Encoding method (BPE) [13]. Another promising approach is to use symbols in the translation process [14]. By converting proper nouns like person’s name to symbols, the model can translate new words [15, 16]. In this paper, we use both BPE with 10,000 subwords and the symbolization process. For the validation of our model, the BLEU score of Eng-Kor and Kor-Eng are 19.0 and 37.45, respectively.

그림 1. 통계 보기 버튼을 포함한 번역 화면
Fig. 1. Translation Interface with Statistics Button in red.

Ⅲ. Generating Personalized Word Lists

The main contribution of this paper is generating personalized word lists in a translation system, aligned with the academic interests of individual OT users. To generate the word lists, we first extracted the words from the source and target text based on white space. We aimed to retrieve specialized words by filtering out 128 stop words [7] and 2,801 general high-frequency vocabulary word families [8]. The high-frequency word families are from the New General Service List (NGSL) developed for English language learners based on a "273 million word subsection of the 2 billion word Cambridge English Corpus" [8].

After generating lists based on the source and target text, we compared these lists to other widely used academic word lists. First, we examined the word overlap with the Academic Word List (AWL) [17] (composed of 570 word families), which is a general English for Academic Purposes (EAP) word list developed from a corpus of 3.5 million words from 28 subject areas across Arts, Commerce, Law, and Science. The NGSL and AWL are expected to cover 90% of the words in academic texts [18]. Second, we examined the word overlap with a science-specific word list (composed of 318 word families) developed from a corpus of academic texts from 14 science-related subject areas [19].

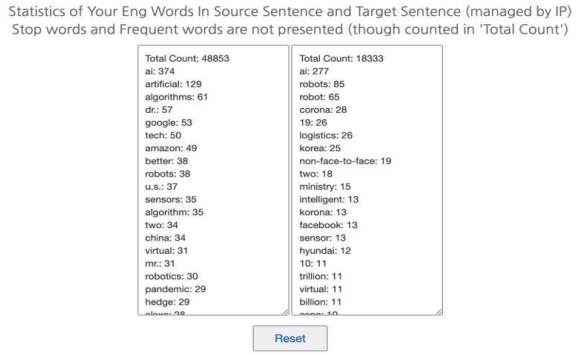

그림 2. 입력과 출력 문장에 대한 단어 리스트 화면
Fig. 2. Word List Interface with source and target lists.

In our NMT system, users can access their word lists by pressing the 'Statistics of Your Translation' button to view the total word count and the 100 most frequent words ranked from the highest to lowest frequency with individual word counts as shown in Figure 1. Users can simultaneously view a list for their source text and a list for their target English text as shown in Figure 2.

Ⅳ. Experiment

To generate a personalized word list for the source text as an example, we inputted 50 English articles with content related to AI from the most popular news websites in the United States. A total of 48,853 words from the 50 articles was inputted into our NMT system, and a list of the 100 most frequent words was retrieved as shown in Table 1. The list included AI-related abbreviations, nouns, adjectives, adverbs, and verbs. In the list, there were common AI terminologies (e.g., algorithm, neural) and trending words (e.g., amazon, pandemic, coronavirus). When compared to the other word lists, only 11 words overlapped with the AWL [17], and one word overlapped with the science-specific list [19].

표 1. 입력 문장에서 가장 자주 나타난 단어 100 리스트
Table 1. 100 Most Frequent Word List (with Individual Count) for Source Text

1–25 Words	26–50 Words	51–75 Words	76–100 Words
ai: 374	facebook: 24	thousands: 14	six: 11
artificial: 129	covid-19: 24	impact: 13	invasive: 11
algorithms: 61	ibm: 24	de: 13	watson: 11
dr.: 57	autonomous: 23	tesla: 13	surveillance: 11
google: 53	facial: 23	pedestrians: 13	ganbreeder: 11
tech: 50	app: 20	warehouse: 13	anchor: 11
amazon: 49	robot: 20	patent: 13	camio: 11
better: 38	media: 20	gallant: 13	infrared: 10
robots: 38	nasdaq: 19	2019: 12	utility: 10
u.s.: 37	a.i.: 19	june: 12	30: 10
sensors: 35	best: 18	millions: 12	2017: 10
algorithm: 35	e: 18	pg: 12	3: 10
two: 34	notco: 18	transmission: 12	billion: 10
china: 34	robotic: 17	ceo: 12	boston: 10
virtual: 31	10: 17	classroom: 12	2015: 10
mr.: 31	electrical: 17	instagram: 12	automated: 10
robotics: 30	institute: 16	resume: 12	intelligent: 10
pandemic: 29	inc: 16	plant-based: 12	inventor: 10
hedge: 29	california: 16	neurons: 12	pentagon: 10
alexa: 28	vivacity: 16	automation: 11	smartphones: 9
million: 27	combat: 15	bullish: 11	5: 9
uk: 26	chinese: 15	2018: 11	1: 9
privacy: 25	yuste: 15	100: 11	internet: 9
coronavirus: 24	kratsios: 15	five: 11	co-founder: 9
neural: 24	2020: 14	healthcare: 11	published: 9

표 2. 출력 문장에서 가장 자주 나타난 100 단어 리스트
Table 2. 100 Most Frequently Word List (with Individual Count) for Target English Text

1-25 Words	26-50 Words	51-75 Words	76-100 Words
ai: 277 robots: 85 robot: 65 corona: 28 19: 26 logistics: 26 korea: 25 non-face-to-face: 19 two: 18 ministry: 15 intelligent: 13 korona: 13 facebook: 13 sensor: 13 hyundai: 12 10: 11 trillion: 11 virtual: 11 billion: 11 espa: 10 defendant: 10 wearable: 9 bullying: 9 baek: 9 warehouse: 9	million: 8 kim: 8 amazon: 8 avatar: 8 kpop: 8 2018: 8 earnest: 8 2025: 8 autonomous: 8 kt: 8 gs: 8 daejeon-si: 8 sensors: 8 b: 8 purification: 8 gyeonggi: 8 five: 7 google: 7 2020: 7 institute: 7 naver: 7 japan: 7 best: 7 productivity: 7 companion: 7	startups: 7 netmarble: 7 posco: 7 jo-kyung: 7 2017: 6 seoul: 6 garbage: 6 actively: 6 sm: 6 four: 6 mobility: 6 assembly: 6 robotics: 6 three: 6 utilization: 6 cctv: 6 algorithm: 6 wonju-si: 6 buddy: 5 convenience: 5 1: 5 20th: 5 coronavirus: 5 conference: 5 strategic: 5	china: 5 pillar: 5 krw: 5 hong: 5 3: 5 distribution: 5 50: 5 optimization: 5 optimal: 5 dumping: 5 inspection: 5 algorithms: 5 precision: 5 accuracy: 5 dispatch: 5 upgrade: 4 accurately: 4 circuit: 4 diagnosis: 4 pandemic: 4 yoo: 4 jang: 4 eight: 4 kda: 4 november: 4

To generate a personalized word list for the target text, we inputted 25 articles in Korean with content related to AI from popular news websites in Korea (e.g., The Kyunghyang Shinmun) and translated the text into English. A total of 18,333 words from the 25 articles were translated, and the statistics for the 100 most frequently translated words were retrieved as shown in Table 2. The list included AI-related abbreviations, nouns, adjectives, adverbs, and verbs. In the list, there were common AI terminologies (e.g., algorithm, robotics), trending words (e.g., pandemic, coronavirus), and context-specific words (e.g., hyundai, krw, k-pop). When compared to the other word lists, only 12 words overlapped with the AWL [17], and 4 words overlapped with the science-specific list [19]. When compared to the source word list, there was an overlap of 31 words with 2 words in a different form and 10 numbers.

The results from the experiment indicate that the personalized word lists generated by our NMT system include highly specific words directly aligned with the interests of individual OT users. These lists contain a variety of useful words such as terminology commonly used in specialized texts and words that provide insight into current area-specific trends. By referring to the lists, students can strategically identify words often used in their academic interest area.

To effectively study the words in the personalized lists, students can begin by reviewing the aspects (e.g., spelling, pronunciation, forms, definition), synonyms, and usage examples of each word using online dictionaries [6]. Then, students can practice retrieving word definitions with paper-based or digital word cards [20]. Finally, students can reinforce their vocabulary knowledge by consistently using the words in their academic English speaking and writing tasks [6].

Ⅴ. Conclusion

In this paper, we proposed using the source and target text from NMT systems to generate personalized word lists to scaffold learning vocabulary. The experiment results demonstrated that our proposed idea provides students with a list of area-specific words that are mostly not found in other widely used academic word lists. This implies that the text in NMT systems can be analyzed to attain valuable insight into increasing specialized English vocabulary knowledge.

For future studies, we plan to generate and evaluate lists with a larger amount of text and for other subject areas. We also plan to generate more meaningful word lists by filtering out certain numbers and proper nouns. Furthermore, we plan to better understand our NMT system by comparing the differences between a word list generated from the source text and a list generated from a back-translation.

References

[1] “Investors - Annual Report”, Naver Corporation, www.navercorp.com/en/investment/annualReport, 2019.
[2] J. Kuczmarski, “A New Look for GoogleTranslate on the Web”, Google, blog.google/products/translate/new-look-google-translate-web/, 2018.
[3] J. Clifford, L. Merschel, J. Munné, "Surveying the landscape: What is the role of machine translation in language learning?", @ tic. revista d'innovació educativa, Vol. 10, pp. 108-121, 2013.
[4] J. R. Jolley, L. Maimone, "Free online machine translation: Use and perceptions by Spanish students and instructors", Learn Languages, Explore Cultures, Transform Lives, pp. 181-200, 2015.
[5] E. M. O’Neill, "Online translator, dictionary, and search engine use among L2 students", CALL-EJ, Vol. 20, No. 1, pp. 154-177, 2019.
[6] A. Coxhead, *Vocabulary and English for Specific Purposes Research: Quantitative and Qualitative Perspectives.* New York, NY, USA: Routledge, 2018.
[7] S, Bleier, “NLTK's List of English Stopwords”, Gist, gist.github.com/sebleier/554280, 2010.
[8] C. Browne, "A new general service list: The better mousetrap

we've been looking for", Vocabulary Learning and Instruction, Vol. 3, No. 1, pp. 1-10, 2014.
[9] K. Cho, B. van Merrienboer, D. Bahdanau, Y. Bengio, "On the Properties of Neural Machine Translation: Encoder-Decoder Approaches", Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation, 2014.
[10] J. Gehring, M. Auli, D. Grangier, D. Yarats, Y. N. Dauphin, "Convolutional Sequence to Sequence Learning", arXiv Preprint, arXiv:1705.03122, 2017.
[11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N.Gomez, L. Kaiser, I. Polosukhin, "Attention Is All You Need", arXiv, 2017.
[12] H. Choi, "Understanding Neural Machine Translation", Communications of the Korean Institute of Information Scientists and Engineers, Vol. 37, No. 2, pp.16-24, 2019.
[13] R. Sennrich, B. Haddow, A. Birch, "Neural machine translation of rare words with subword units", 54th Annual Meeting of the Association for Computational Linguistics, 2016.
[14] H. Choi, K. Cho, and Y. Bengio, "Context-dependent word representation for neural machine translation", Computer Speech & Language, Vol. 45, p. 149-160, 2017.
[15] C. Kang, Y. Ro, J. Kim, H. Choi, "Symbolizing Numbers to Improve Neural Machine Translation", Journal of Digital Contents Society, Vol. 19, No. 6,p.1161-1167, 2018.
[16] J. Nam, M. Kim, H. Jeong, H. Choi, "Kor-Eng Neural Machine Translation System Using Proper Noun Dictionary" KCC, 2020.
[17] A. Coxhead, "A new academic word list", TESOL Quarterly, Vol. 34, No. 2, pp. 213-238, 2000.
[18] S. Smith, "What Is the AWL?", EAPFoundation.com, www.eapfoundation.com/vocab/academic/awllists/, 2020.
[19] A. Coxhead, D. Hirsch, "A pilot science-specific word list", Revue française de linguistique appliquée, Vol. 12, No. 2,pp. 65-78, 2007.
[20] I. S. P. Nation, "Making and using word lists for language learning and testing", John Benjamins Publishing Company, 2016.

우진하(Jin-Ha Woo)

2009: Dept. of Second Language Studies, University of Hawaii at Manoa (B.A.)
2012: Dept. of Second Language Studies, University of Hawaii at Manoa (M.A.)
2011년~2012년: University of Hawaii at Manoa (English Language Instructor)
2012년~2014년: Hitotsubashi University (Adjunct Assistant Professor)
2014년~2020년: Dept. of Language Education, Handong Global University (Visiting Professor)
※관심분야 : Educational Technology, Technology Integration Professional Development

최희열(Heeyoul Choi)

2010: Dept. of Computer Science and Engineering, Texas A&M University (Ph.D)
2010~2011: Indiana University (PostDoc)
2015~2016: University of Montreal (Visiting Researcher)
1998년 ~ 2001년: OromInfo (Programmer)
2011년~2016년: 삼성전자 종합기술원 (Research Staff Member)
2016년~2020년: 한동대학교 전산전자공학부 조교수
2020년~현재: 한동대학교 전산전자공학부 부교수
※ 관심분야 : Machine Learning, Deep Learning, Artificial Intelligence, Neural Machine Translation