A Study on Usage Frequency of Translated English Phrase Using Google Crawling

Kyuseok Kim*, Hyunno Lee**, Jisoo Lim**, Sungmin Lee**
*Professor, Dept. of Data Convergence Software, Bundang Convergence Technology Campus, Korea Polytechnics
**Dept. of Data Convergence Software, Bundang Convergence Technology Campus, Korea Polytechnics
kyuseokkim@kopo.ac.kr, soulnbeat01@gmail.com, jishuya3015@gmail.com, lssmm1230@naver.com

Abstract

People have studied English using online English dictionaries when they looked for the meaning of English words or the example sentences. These days, as the AI technologies such as machine learning have been developing, documents can be translated in real time with Kakao, Papago, Google translators and so on. But, there has still been some problems with the accuracy of translation. The AI secretaries can be used for real-time interpreting, so this kind of systems are being used to translate such the web pages, papers into Korean. In this paper, we researched on the usage frequency of the combined English phrases from dictionaries by analyzing the number of the searched results on Google. With the result of this paper, we expect to help the people to use more English fluently.

1. Introduction

In the 4th industrial revolution, we get to the foreign culture such as news, documents, TV shows so closer than ever because we can quickly connect to the Internet anytime, anywhere. In particular, we can get a lot of useful information faster than ever with the rapid development of machine translation technology by AI[1][2]. In the previous research, they studied on the use of machine translator such as Kakao, Papago, Google translator and its effect on high school students' English writing[3].

However, the accuracy of the machine translating Korean into English phrases was sometimes low because there's some problem like when the translator can't recognize the object or verb properly[4]. Moreover, we could get struggling with choosing the proper English words when we try translating Korean into English sentences using English dictionaries.

In this reason, we propose a method to compose English sentences using English dictionaries by calculating the usage frequency of translated English phrase analyzing the number of the searched results on Google.

2. Experimental Environment

We gathered the experimental data from open vocabulary book of Daum Dictionary. We collect the example sentences consisting of verbs and objects from "[회화] 가장많이쓰이는동사-랭킹 Top1000[01]" in the vocabulary book. Among the 1000 of phrases, we randomly chose 50 ones as Figure 1[5].

```
text = [
    "선물을 받다", "시설을 사용하다", "지갑을 발견하다",
    "편을 놓다", "도움이 필요하지 않습니다 ", "직업을 물어보다",
    "친구를 초대했다", "음식을 제공하다", "손잡이를 돌리다",
    "사건이 뒤따르다", "시골로 향하다", "꽃병을 높혀 놓다",
    "수업료를 내다", "목소리가 들리다", "불순물을 포함하다",
    "호텔에서 만나자", "멈춰서서 싸우다", "일거리를 제공하다",
    "뇌물을 주다", "마루를 닦아야 되겠다", "성격을 바꾸다",
    "세탁물을 보내다", "문제를 해결하다", "아이를 다루다",
    "능력을 발전시키다", "사건을 묘사하다", "컨펀에 도착하다",
    "해군 기지로 만들다", "논쟁에 말려들다", "식비에 돈을 들이다",
    "여행에서 돌아오다", "소포를 부치다", "비가 그치다",
    "편지를 받다", "특별상을 받다", "복장을 바꾸다",
    "주운 날씨가 계속되다", "직장을 잃다", "배에서 떨어지다",
    "낯선사람이 나타났다", "노래가 들리다", "접시가 운반되다",
    "사과를 갖다줘", "정각에 출발하다", "표를 보여주다",
    "텔레비전을 부착하다", "편지를 보존하다", "많은 돈을 잃어버리다",
    "제안을 충분히 검토했다", "물건을 어깨에 메고 나르다"
]
```

(Figure 1) Experimental Data

To get the result, we composed a program using python. To operate the procedures automatically, we used the python module called “Selenium[6].” Additionally, we used the “konlpy” module to separate the object and verb[7].

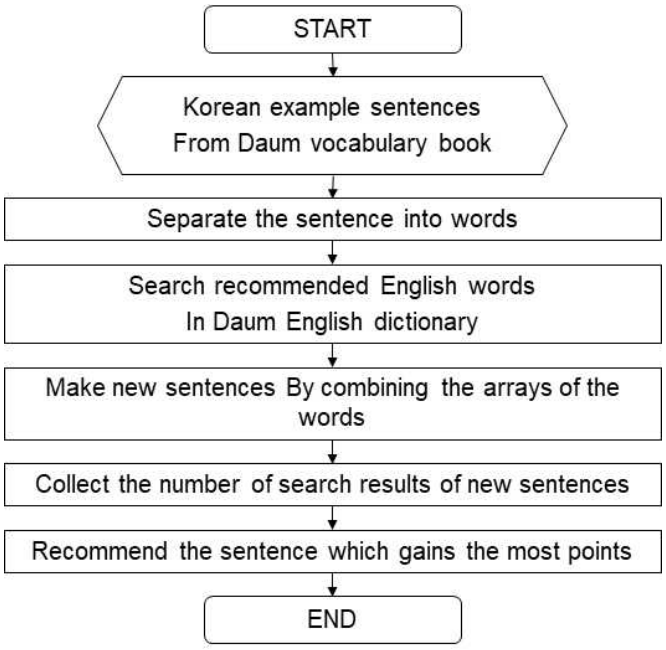
To measure the correlation between the index and the proportion variables, we used the Pearson Correlation Coefficient as shown formula (1)[9]. This has a value between －1 and 1. The positive value is positive linear, the negative value is negative linear. The larger the absolute value, the stronger the relationship between the variables.

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} \quad (1)$$

cov : covariance
σ_X : standard deviation of X
σ_Y : standard deviation of Y

In some case, the verb and the object are not completely separated. For example, the phrase "제안을 충분히 검토했다" is divided into "제안" as the object and "검토충분하다" as the verb.

The flow chart for this operation shows as Figure 2. At first, the “해군기지” and “만들다” of Korean phrases are translated into English on Daum dictionary[8]. On Daum dictionary, it usually recommends multiple English words. One phrase usually contains one object and one verb. Therefore, the number of combinations from the recommended English words can be calculated as the number of objects multiplied by the number of verbs. Then, these combined English phrases are searched on Google and the number of searched results is also crawled.



(Figure 2) Flow Chart for Operation

3. Experimental Result

Take for instance, using “해군 기지로 만들다” of Korean phrases, we could get the proposed “build naval base” of English phrase. And the number of combined English phrases is shown as Figure 3.

```
문장분해: ['해군기지', '만들다']
명사크롤링 단어: ['naval base']
동사크롤링 단어: ['make', 'create', 'build', 'form', 'write']
검색문장: "make naval base" /검색결과 갯수: 1240
검색문장: "create naval base" /검색결과 갯수: 743
검색문장: "build naval base" /검색결과 갯수: 14200
검색문장: "form naval base" /검색결과 갯수: 616
검색문장: "write naval base" /검색결과 갯수: 3
```

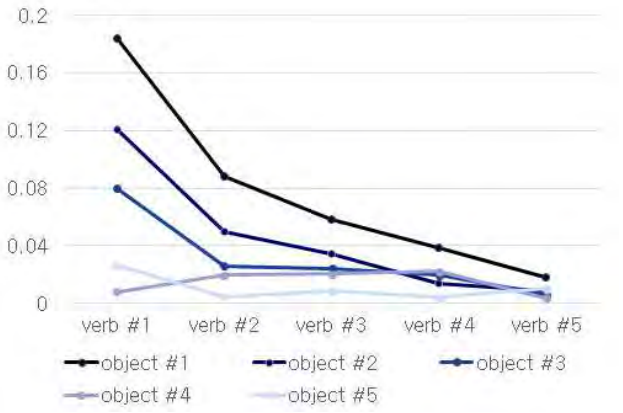
(Figure 3) Combined English Phrases

The average numbers of the recommended words for object and verb from Daum dictionary are 3.40 and 3.56 for each. And we combined the words by order and searched for the combined phrase on Google. The numbers of the searched results are shown as Table 1. The proportion of the combined phrase depends on the distance between the object and the verb indexes.

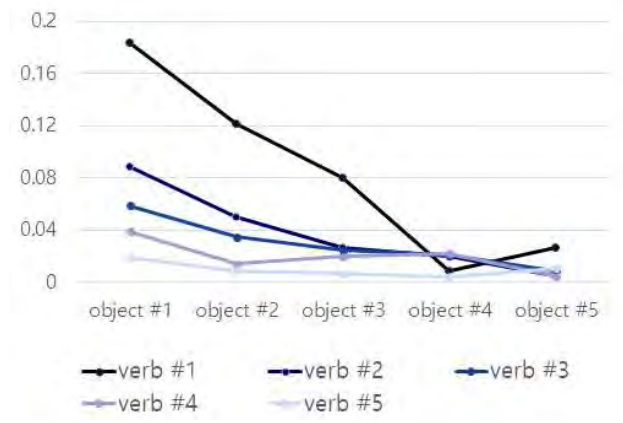
<Table 1> Proportion Search Count of Combination Words

verb / object	#1	#2	#3	#4	#5
#1	0.1835	0.1210	0.0797	0.0083	0.0264
#2	0.0882	0.0496	0.0261	0.0196	0.0048
#3	0.0585	0.0346	0.0242	0.0204	0.0085
#4	0.0389	0.0138	0.0195	0.0221	0.0044
#5	0.0184	0.0088	0.0062	0.0041	0.0104

The proportions based on object are shown as Figure 4 and those based on verb are shown as Figure 5. According to them, the proportions usually decrease.



(Figure 4) The Proportions Based on Object



(Figure 5) The Proportions Based on Verb

To measure the correlations, the pearson correlation coefficient values were calculated as Table 2 and Table 3. Table 2 and Table 3 show the values according to the Row(Object), Column(Verb) respectively. The absolute values of them are usually close to 1, so they are usually strongly negative linear.

<Table 2> Pearson Correlation Coefficients by the Row Index

	Pearson's r Value
#1	-0.947
#2	-0.956
#3	-0.959
#4	-0.757
#5	-0.596

<Table 3> Pearson Correlation Coefficients by the Column Index

	#1	#2	#3	#4	#5
Pearson's r Value	-0.928	-0.909	-0.860	-0.115	-0.568

4. Conclusion

The aim of this paper is to get the most commonly used phrases by English native speakers. So, we proposed a method to compose English phrases which are the most commonly used.

With combining the words and calculating the results, the combinations with the first recommended object and the first recommended verb are usually the most frequently used on Google. The highest point of the combination words is 0.1835. This indicates that the combination of the first recommended object and the first recommended verb is the most frequently used phrase by calculating the number of the searched results on Google. Moreover, the correlation between the recommended index and the proportion is usually negative linear. Thus, this indicates that the combination of the first recommended object and the first recommended verb is usually more used phrase than the others.

In the future researches, we could propose a method to make English sentences which are more proper to the context. Moreover, it could help the real-time interpreters to translate languages more fluently.

References

[1] Y. Yue, J. Lee, “Translation of Verb Tense Marks in Korean Chinese and Korean English Machine Translation”, Journal of Korean Language Education, Vol. 14, No. 2, pp.36-60, 2019.07

[2] S. H. Lee, S. H. Kim, “Pre-editing Rules to Enhance Output Quality of Machine Translation : English-Korean and Korean-English”, The Journal of Translation Studies, Vol. 19, No. 5, pp.121-154, 2018.12

[3] Y. J. Lee, D. J. Lee, “A Study on the Use of Machine Translator and its Effect on High School Students’ English Writing”, Journal of the Korean English Education Society, Vol. 19, No. 2, pp.159 ~180, 2020.05

[4] O. S. Park, “Error Analysis According to the Typological Characteristics of Source Text in Korean-English Machine Translation”, The Journal of Society for Humanities Studies in East Asia, Vol. 41, pp.155-183, 2017.12

[5] https://wordbook.daum.net/open/wordbook/list.do?dic_type=endic

[6] https://www.selenium.dev/

[7] https://konlpy-ko.readthedocs.io/ko/v0.4.3/

[8] https://dic.daum.net/index.do?dic=eng

[9] S. H. Kang, I. S. Jeong, H. S. Lim, “A Feature Set Selection Approach Based on Pearson Correlation Coefficient for Real Time Attack Detection”, Journal of Convergence Security, Vol. 18, No. 5, pp.59-66, 2018.12