채식주의자: 랭귀지 모델 접근

김재준[0], 권준혁, 김유래, 박명관, 송상헌

동국대학교, 인천대학교

kjj8286@naver.com, sanghoun@inu.ac.kr

A Language Model Approach to "The Vegetarian"

Jaejun Kim[0], Junhyeok Kwon, Yoolae Kim, Myung-Kwan Park, and Sanghoun Song

Dongguk University, Incheon National University

요 약

This paper is to broaden the possible spectrums of analyzing the Korean-written novel "The Vegetarian" by using the computational linguistics program. Through the use of language model, which was usually used in bi-gram analysis in corpus linguistics, to the International Man Booker award winning novel, the characteristics of "The Vegetarian" is investigated by comparing it to the English-written novel "A Little Life".

주제어: Language Model, N-gram Analysis, The Vegetarian

1. Introduction

In this research, the comparison between the 2016 International Man Booker award winning "The Vegetarian" written by Han Kang and the 2015 Man Booker award nominee "A Little Life" written by Hanya Yanagihara is conducted by using the language model analysis. Using "A Little Life" as the reference novel, the translation of the original text, "The Vegetarian" is thoroughly investigated.

As the language models are now being applied in many different kinds of linguistic areas, this study aims to contribute to the area of translation. So far, the researches using the language models compared only the limited expressions between the novels, whereas in this research, comprehensive analysis is conducted. The reference novel and the target novel are different in that their original languages differ, one being English and the other being Korean. Thus, comparing these books can be meaningful in terms of translation. Pattern of sentences or phrases from each novel can also be investigated through the application of the language model.

There are many kinds of language models such as N-gram model, structured model, and class-based model. Among these forms of language modeling, N-gram model will play a pivotal role as a basic foundation. In other words, deep and various analyses of "The Vegetarian" is carefully observed.

2. Background

As mentioned above, when interpreting the literature works, researches using various types of analyses compared only the limited expressions, a word level. Thus, in this research, comprehensive analysis is applied to observe the distinctive patterns within the novel. Since "The Vegetarian" is the first Korean novel that was awarded for the International 2016 Man Booker award, extensive analysis through distant reading approach is necessary in order to find its special aspects.

There are three main reasons for choosing "The Little Life" as a reference novel. First, the general plots between the novel are similar. Both novels have the contents of psychic trauma in

the past, desire, and suicide of the main character. Second, the text itself is about 2.5 times larger than "The Vegetarian". Thus, the larger size of the text can provide excellent standard for being the reference text. Third, "A Little Life" is originally written in English. Since "The Vegetarian" is originally written in Korean, there may be certain characteristics of translated novels. In that way, "A Little Life" can display a criterion of the originally English-written text. For these reasons, "A Little Life" can play an important role as a reference text for "The Vegetarian".

3. Previous Analysis (Roh (2017))

According to Moretti (2005), reading literature writings are divided by two big categories, distant reading and close reading. Other than the traditional way of reading literatures, Moretti (Ibid.) proposed to visualize the text by counting texts and making graphs and maps, as a main characteristic of distant reading. In distant reading, quantified results, which cannot be obtained by close reading, are presented by analyzing the text as a whole.

By implementing distant reading of "The Vegetarian", Roh (2017) divided the text into three different parts where the viewpoint slightly changes in each part. Rho (Ibid.) mainly investigated frequent words of both a whole text and each parts such as (1) and (2).

(1) Frequent Word List (Entire Text)

	Word	Frequency		Word	Frequency
1	like	208	7	face	122
2	just	190	8	eyes	120
3	time	190	9	he'd	120
4	yeonghye	160	10	she'd	116
5	wife	141	11	body	100
6	inhye	125	12	It's	82

(2) Frequent Word List (Part 1)

	Word	Frequency
1	wife	105
2	just	51
3	meet	46
4	time	40
5	life	39
6	face	38

As can be seen from (1) and (2), her work mainly focuses on the unigram analysis. It is noticeable that the frequent word lists slightly changes depending on which part is analyzed.

However, there are several shortcomings to this analysis. First, the POS (Part Of Speech)-tagging was not implemented from the beginning. Words such as '*like*' are used as different POS, depending on the sentences. However, in Roh (Ibid.)'s analysis, '*like*' was investigated according to its different POS usages after already being counted as one as in (1). Therefore, the accurate count of the word '*like*' is somehow confounded. Second, in Roh (Ibid.)'s work, neither BOS (Begin Of the Sentence) nor EOS (End Of the Sentence) were considered. Third, similar to the second reason, Rho (Ibid.) excluded exclamation marks and question marks. The last two reasons are all connected with each other. Roh (Ibid.) did not distinguish the sentences because the main focus was on the unigrams. However, when it comes to n-grams other than just the unigrams, considering the BOS and the EOS is highly important. The bigrams should be limited to sentences because the two consecutive words from different sentences must not be counted as bigram words. Thus, dividing the text as sentence by sentence is highly significant.

4. Methodology

4.1 Language Modeling

Language model refers to the statistical probability of sequential words combination. Among many types of language model, three representative models exist in language processing which include

n-gram model, class-based model, and structured model. Among these types, n-gram model will play a pivotal role as a basic foundation, as mentioned earlier. In this way, n-gram probabilities can help find specific word patterns for "The Vegetarian". Our research used SRILM (Stanford Research Institute Language Modeling) toolkit in applying n-gram language model.

4.2 Procedures

Before applying language model to the text, the text must be preprocessed in order to prevent erroneous results and interpretations. First, tokenization is the start of the preprocessing the given text. Tokenization includes lowering all the capitalized letters, tagging the POS of the words, and removing hyphens. In this way, the words from the text are equally counted and their information is also properly calculated. POS-tagging was conducted by using Standard POS tagger. By implementing these procedures, the text is ready for the language model analysis.

After tokenizing the text, computing the language model is the next step. Thus, by running language model, the frequencies and probabilities of the words can be investigated. One possible problem can occur when counting the frequencies of the words. Since the size of the text can differ from each other, the absolute frequencies cannot be the suitable indicator for the analysis. In language model, this problem is removed because the language model itself reflects the relative frequencies of the words. Thus, it is possible for the researchers to compare frequencies between texts. As a final procedure, smoothing is required in order to avoid possible underflow problem. As its tool, extended interpolated Kneyser-Ney is applied.

5. N-Gram Analysis

As a outcome of the language model, it is shown as an ARPA format such as (3) and (4). APRA format allows us to briefly analyze the result of the language model in a certain form. In APRA format, it is divided into two main parts. Left column show the basic form of the column on the right. The right column describes the actual results of the text. The right column is composed of the number of n-grams at the top. On the bottom, the numbers present the probability of the certain words. As mentioned earlier, the probability is relatively calculated. To the right, the words and their POS-tagging are displayed. Thus, this format allows us to analyze the data in a more efficient and convenient way.

(3) ARPA Format of "The Vegetarian"

/data/	/data(type)/	
ngram 1=n1	ngram 1=6741	
ngram 2=n2	ngram 2=30121	
…	ngram 3=4572	
ngram N=Nn	ngram 4=1829	
/1-grams:	ngram 5=745	
P w [bow]	/1-grams:	
…	-2.435441	like_in
/2-grams:	-3.82404	like_vb
P w1 w2 [bow]	-4.23738	like_vbp
…	…	
/N-grams:	/3-grams:	
P w1 … wN	-0.1808866	there_ex 'll_md be_vb
…	/4-grams:	
/end/	-0.3314916	the_dt nurses_nns '_pos room_nn
	/end/	

In (3), as can be seen under *data(type)*, there is a reduction in the frequently used word combinations from *2-grams* to *5-grams*. It means that since the language model distinguished all the sentences, the number of n-grams decreases depending on the number of the n-grams. Another notable aspect of (3) is that the different POS-tagged usage of *like* is presented in detail under the different POS categories. The different POS of *like* such as *preposition, base form verb,* and *non-4rd person singular present verb* are presented. The prepositional usage of *like* shows the highest probability among other usages, followed by verb usages. On top of that, in 3-grams and 4-grams, the result shows that the language model works perfectly, depending on the given text. Especially in 4-grams, one of the most frequently used word combinations include the word *nurse* because the main content of "The Vegetarian" involves hospital. Thus, the sentences that the language

model calculated are affected by the contents.

(4) ARPA Format of "A Little Life"

```
/data/                  /data(type)/
ngram 1=n1              ngram 1=17638
ngram 2=n2              ngram 2=110999
...                     ngram 3=31373
ngram N=Nn              ngram 4=18793
/1-grams:               ngram 5=9922
P w [bow]               /1-grams:
...                     -2.769982       like_in
/2-grams:               -4.023671       like_vb
P w1 w2 [bow]           -4.304081       like_vbp
...                     -4.78211        like_jj
/N-grams:               ...
P w1 ... wN             /3-grams:
...                     -0.5745651      how_wrb 'd_md you_prp
                        ...
/end/                   /end/
```

In terms of (4), the number of *data(type)* of "A Little Life" is considerably higher than "The Vegetarian", as it is chosen for the reference text. Thus, the number of n-grams reflect the actual volume of the text. Different from (3), it is noticeable to take a look at the variety usage of the word *like*. In this novel, which is originally written in English, another POS of *like* is included. In here, usage of adjective of *like* is added, but its usage is the lowest among *like*. The probability of seeing *like* from both novels displays similar outcome. The most frequently used POS of *like* is the same. Thus, even though the detailed usage of *like* is different in that "A Little Life" used *like* as an adjective, the overall trend of using the word is almost similar.

(5) Bi-grams of "The Vegetarian" and "A Little Life"

The Vegetarian	Probability	A Little Life	Probability
such_pdt a_dt	-0.07123	wo_md n't_rb	-0.02217
able_jj to_to	-0.07909	ca_md n't_rb	-0.04107
unable_jj to_to	-0.07909	able_jj to_to	-0.05225
wo_md n't_rb	-0.08060	lack_nn of_in	-0.05487
trying_vbg to_to	-0.08276	supposed_vbn to_to	-0.06016
want_vbp to_to	-0.10461	hundreds_nns of_in	-0.06022
began_vbd to_to	-0.10617	lots_nns of_in	-0.06022
ca_md n't_rb	-0.10668	kinds_nns of_in	-0.07481
kind_nn of_in	-0.11140	unable_jj to_to	-0.08094
does_vbz n't_rb	-0.11628	version_nn of_in	-0.09753

On top of the ARPA format of both texts, simple bi-gram comparison depending on the probabilities display another similarity as in (5). Since there are two many word combinations in both texts, (5) only briefly lists the words. Among the top ten bi-grams that are likely to appear, both texts had quite a lot of overlapping word combinations. From the top ten bi-grams, four combinations were used in both text with a high probability. About 40 percent of the used word combinations are overlapped.

6. Conclusion

As a conclusion, there is no distinctive characteristics of "The Vegetarian" itself. Although "A Little Life" can show more diverse usages of words, the overall usage of the context is similar. The reason for that is because the originally Korean-written novel "The Vegetarian" and English-written novel "A Little Life" did not display differences in word sequence usages. Even though "The Vegetarian" is translated into English after written in Korean, the translation was focused on conveying the meaning of the original sentence, not on sentence-to- sentence correspondence. From this point of view, we observed that the sketchy features of the novels can be investigated through the means of the language model. With the help of the language model approach, researchers can distinguish whether the novel is written in a English-like way or not in a way.

This research is understandably not perfect in terms of choosing the reference text. In order to develop this analysis of applying language model to the literature works, the reference text needs to be bigger than "A Little Life" in order to set the genuine standards. Thus, this research can be supplemented with the help of the bigger reference text.

References

[1] Moretti, F. (2005). *Graphs, maps, trees: abstract models for a literary history*. Verso.

[2] 노은주. (2017). 문학 작품 "멀리서 읽기":한강의<채식주의자> 번역본 텍스트 분석. *언어와 언어학*, *74*, 75-104