Solving English Questions through Applying Collective Intelligence

Dongjin Choi[1], Myunggwon Hwang[2], Byeongkyu Ko[1], and Pankoo Kim[1]

[1] Dept. Of Computer Engineering Chosun University, Gwangju, South Korea
[2] Korea Institute of Science and Technology Information, Daejeon, South Korea
{Dongjin.Choi84,rhqudrb135}@gmail.com,
mgh@kisti.re.kr, pkkim@chosun.ac.kr

Abstract. Many researchers have been using n-gram statistics which is providing statistical information about cohesion among words to extract semantic information in web documents. Also, the n-gram has been applied in spell checking system, prediction of user interest and so on. This paper is a fundamental research to estimate lexical cohesion in documents using trigram, 4gram and 5gram offered by Google. The main purpose of this paper is estimating possibilities of Google n-gram using TOEIC question data sets.

Keywords: N-gram, Natural Language Processing, Semantics.

1 Introduction

N-gram has been applied to various fields for information processing to estimate semantic information in web documents and to analyze semantic relation between words. The n-gram information is a statistical data set collected and extracted from huge web document sets through analyzing frequency between adjacent words. The n-gram consists of bigram, trigram, 4gram and 5gram and the trigram is the most common data of the n-gram. The n-gram is based on probabilities of an adjacent words occurrence. For this reason, the n-gram can be a fundamental data set in natural language processing and word recommendation field. For instance, there is a difference and similarity between English and Chinese language. [1] compared how English statistically far apart from Chinese and how close they are using n-gram information. Also, the n-gram was applied to text processing [2] and user demands forecasts system [3] and so on. This paper focused on text processing using n-gram data sets provided by Google. Google n-gram contains approximately 1.1 billion words that occurred more than 40 times in documents set what Google had in 2006. Google has been using this n-gram data to query recommendation system. Moreover it is applied to speech recognition [4] and word recommendation text editor [5]. This paper estimates usability of each Google n-gram using one of the representative English language tests in Korea TOEIC as an experimental data set. The differences between each n-gram recall and precision rate will be presented.

The reminder of the paper is organized as follows. Section II describes what Google n-gram and TOEIC are and related works of this paper. A method to apply

J.J. Park, L.T. Yang, and C. Lee (Eds.): FutureTech 2011, Part I, CCIS 184, pp. 37–46, 2011.

Google n-gram data to TOEIC is described in Section III with examples. In Section IV, experimental results are given based on our forecasts system. Finally, Section V concludes with a summary of the results and gives challenging problems.

2 Related Works

2.1 Fundamental Data

Google N-gram. There is no doubt that Google is the most famous and representative search engine in the world. Google contains huge web pages concerning diverse fields such as news, conference and government information and so on. Lots of researcher and public users are using Google search engine to resolve their curiosities. Google has own strong retrieval strategy so that they can save lots of useful web documents. Google grasped n-gram data sets from various web documents and these n-gram data sets are provided by LDC (Linguistic Data Consortium)[1]. The Google n-gram data sets consist of approximately 1.1 billion words occurred more than 40 times in web documents sets (about one trillion word terms). For this reason, we expect the reliability of results using Google n-gram is high. Table 1 shows statistics of each Google n-gram.

Table 1. The number of tokens of each n-gram

N-grams	Number of Tokens
Unigram	13,588,391
Bigram	314,843,401
Trigram	977,069,902
4gram	1,313,818,354
5gram	1,176,470,663
Total token	1,024,908,267,229
Total sentences	95,119,665,584

TOEIC. For more than 30 years, the TOEIC test has set the standard for assessing English-language listening and reading skills needed in the workplace. More than 10,000 organizations in 120 countries throughout the world trust the TOEIC test to determine who has the English language skill to succeed in the global workplace. Because of this fact, most of companies and universities in Korea are using TOEIC as a criteria to determine his (her) English ability. It is divided in Section I (Listening) with 4 kinds of parts and Section II (Reading) with 3 kinds of parts. The fifth part is a multiple-choice assessment that has incomplete sentences with four different possible answers. We use the fifth part as an experimental data set provided by one of famous English education web site named Hackers[2] to estimate usability of Google n-gram. Following table 2 gives examples of fifth part in TOEIC test.

[1] http://www.ldc.upenn.edu
[2] http://www.hackers.co.kr

Table 2. Examples of fifth part in TOEIC test

No.	Questions
1	According to the recent survey by the Transport Committee, the subway system is considered the most ------ means of transportation. (A) preference (B) preferred (C) preferring (D) prefer
2	The state commissioner said that signs ----- the exit route throughout the company premises should be posted. (A) indicating (B) indication (C) indicate (D) indicates
3	Meyers Shoes.com is well known for its widest ----- of women's and men's shoes and accessories. (A) display (B) selection (C) placement (D) position
…	…

This paper contains basic research for analysing lexical structure in web documents using n-gram data frequency. To test usability of Google n-gram, we made simple forecast system that chooses which word can be the best answer to the blank using trigram, 4gram and 5gram in Google n-gram. Therefore, the usability and precision rate will be evaluated.

2.2 Similar Works

N-gram is a collective intelligence frequency data among adjacent words so it has been applied to NLP (Natural Language Processing), Parsing and Terminology Extraction and so on. The n-gram can be described in two different ways. First is the gram by each character in sentences and second is gram by each word. For example, when we have a sentence "The best thing about the future is that it comes one day at a time" by Abraham Lincoln, each n-gram can be extracted as follows using first and second method.

→ The first method
trigram: {The, heb, ebe, bes, est, …} #50
4gram: {Theb, hebe, ebes, best, estt, …} #49
5gram: {Thebe, hebes, ebest, bestt, estth, …} #48
→ The second method
trigram: {The best thing, best thing about, thing about the, …} #13
4gram: {The best thing about, best thing about the, thing about the future, …} #12
5gram: {The best thing about the, best thing about the future, thing about the future is, …} #11

Because the total size of n-gram based on first method is bigger than second one, it takes more time to retrieve and calculate the n-gram data sets [6]. Also, it is not easy to match human spoken language with n-gram data because most of the n-gram data sets are written word from [6]. For example, onomatopoeia such as "um", "uh" is frequently used in spoken English even though it has no meaning but not in written English. Therefore, the n-gram data extracted from spoken and written English are different even they have the same meanings. To reduce the differences between spoken and written English, [6] suggests a method to merge 4gram data sets in spoken

English to trigram data set in written English. Besides, n-gram is applied to estimate lexical structure of documents to find semantics. [7] proposed a method to decide which words such as pronoun and preposition are suitable in sentences using Google n-gram data sets. Additionally, n-gram has been used to estimate noun compounds in specified domain documents to determine what keyword it is. The author of [8] presented a method to analyse which glossary term will be precisely represent documents in Grolier[3] data sets collected by Mark Lauer based on bigram data frequencies. The biggest obstacle of Google n-gram is the size of data. The fourth and fifth grams are nearly 23GB each but trigram is 15GB. Lots of researches are based on trigram because bigram is too short to find semantics and fourth and fifth are too big to satisfy costing time. To overcome this limitation, [9] gave an idea to modify the threshold value which is a criteria when extract n-gram frequencies. As we can see from above researches, n-gram model has been dynamically used in various fields because it is reliable data set.

3 A Method to Apply Google n-Gram

This section describes a method to apply Google n-gram data sets to TOEIC test questions. There are 4 kinds of possible answers in fifth part of TOEIC. Questions are incomplete sentences with a blank. Therefore, there are at least four possible candidates for correct answer when using 5gram if the blank occurred at the front or the end. If the blank placed at the middle of sentences, the possible candidates for correct answer are up to twenty using 5gram. The number of candidates in a sentence using each n-gram was followed by given formula (1). Following formula (2) indicates total number of candidates in a sentence with four kinds of possible answer.

$$T = 2 \text{ x } n - 1 \text{ , } NG = T - n + 1 \text{ .} \tag{1}$$

where, T is total number of terms inputted in the system including blank, n is depth of n-gram and NG is total number of constructed n-gram.

$$TNG = NG \text{ x } 4 \text{ .} \tag{2}$$

where, TNG is the total number of constructed n-gram including 4 kinds of possible answers.

The position of the blank in sentences and depth of n-gram determine the number of candidates. Following table 3 shows examples of possible n-gram determined by position of the blank.

Table 4 gives examples of fifth part in TOEIC test with frequencies from Google 5gram data. The system compares the 5gram from given sentence with 5gram in Google. If these 5grams are matched, the word which has the highest frequency rate will be determined as an answer. This is based on the fact that the most frequently used sentences have the highest probability for correct answer. The system sum the frequencies of each possible answers and the highest one decided as a correct answer. This paper contains the basic research to assess usability of Google n-gram. The evaluation of recall and precision rate are based on n-gram frequencies provided by Google.

[3] http://en.wikipedia.org/wiki/Grolier

Table 3. Example of candidates n-gram determined by position of the blank

Blank at the front	
Example	----- the store is understaffed right now, ~ (A) Although (B) Yet (C) Meanwhile (D) But
5gram	Although the store is understaffed, Yet the store is understaffed, Meanwhile the store is understaffed, But the store is understaffed
	$NG = T - n + 1 = 5 - 5 + 1 = 1$, $TNG = NG$ x 4 = 4
Blank at the end	
Example	~ after a colleague said the slides would be too -----. (A) distract (B) distracted (C) distractedly (D) distracting
4gram	would be too distract, would be too distracted, would be too distractedly, would be too distracting
	$NG = T - n + 1 = 4 - 4 + 1 = 1$, $TNG = NG$ x 4 = 4
Blank at the middle	
Example	~ within 90 days and ----- by the barcode on the box. (A) altered (B) adjusted (C) accepted (D) accompanied
Trigram	days and altered, days and adjusted, days and accepted, day and accompanied, and altered by, and adjusted by, and accepted by, and accompanied by, altered by the, adjusted by the, accepted by the, accompanied by the
	$NG = T - n + 1 = 5 - 3 + 1 = 3$, $TNG = NG$ x 4 = 12

Table 4. Example of fifth part in TOEIC and its 5grams

Question & Answer		**~ in order to inform ( ) about the purpose of ~** **(A) themselves (B) them (C) that (D) it**	
5gram candidates	themselves	- in order to inform themselves	170
		- order to inform themselves about	68
		- to inform themselves about the	1858
	them	- in order to inform them	2825
		- order to inform them about	720
		- to inform them about the	8980
		- inform them about the purpose	47
		- them about the purpose of	302
	that	- in order to inform that	118
	it	- in order to inform it	158
		- to inform it about the	467
		$NG = 5$, $TNG = 20$	
Sum		themselves: 2096 that: 118	them: 12874 it: 652

4 Evaluation and Results

The usability, efficiency and precision of Google n-gram are compared and evaluated with each n-gram through experiments in this section. For the evaluation, the n-gram

which had special character has been removed so total sizes of trigram, 4gram and 5gram data sets are approximately 24GB, 24GB and 13GB respectively. The TOEIC test sentences for the evaluation were provided by Hackers TOEIC which is one of popular English educational web sites. 200 questions were randomly chosen from the Hacker TOEIC and were compared with Google n-gram data sets. We made a simple system to automatically determine which word will be correct answer for the question shown in figure 1. This system needs two kinds of inputs. First one is an incomplete question sentence including blank and second one is four kinds of possible answers. The system compares every possible 5gram candidates extracted from question sentence with Google 5gram data and saves total frequency of each word if they are matched. Eventually, the system found eleven matched 5grams between question data and Google 5gram data shown in figure 1. The frequencies of each word are 2096, 12874, 229, 625. The system chose 'them' as a correct answer due to its highest frequency.

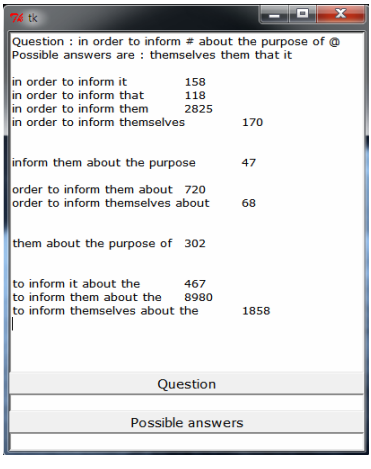

Fig. 1. Example of the system

Following table 5 indicates frequencies and summations of each word when system finds the answer using Google trigram and 4gram. According to our system, 'procedure' and 'coordinating' were chosen as answer and these are same with precise answers.

There are two types of question in TOEIC part five. First is a vocabulary question which stems of given example words are not same. Second is grammar question which stems of given words are the same. Moreover, in natural English language, it is hard to count the number of way to express English. In order to test whether our system satisfies in dynamic English expression or not, we need to compare recall, precision and F1 rate of vocabulary and grammar question based on following formula (3) and (4). Table 6 gives examples of these two types of question data.

Table 5. Example of fifth part in TOEIC and its 5grams

Using 4gram		
Question & Answer	... follow the standard # if they are @ ...	
	(A) procedures (B)developments	
	(C) categories (D) qualifications	
4gram candidates	- qualifications if they are	216
	- categories if they are	253
	- developments if they are	190
	- follow the standard procedures	575
	- procedures if they are	880
Sum	**procedures: 1655**	developments: 190
	categories: 253	qualifications: 216
Using trigram		
Question & Answer	... charge of # the projects @ ...	
	(A) collaborating (B)intending	
	(C) pending (D) coordinating	
4gram candidates	- of coordinating the	25477
	- of intending the	112
	- charge of collaborating	117
	- charge of coordinating	10708
	- charge of intending	59
	- of pending the	91
	- coordinating the projects	412
Sum	collaborating: 117	intending: 171
	pending: 91	**coordinating: 36597**

Table 6. Two types of Question in part five

A type of the vocabulary question
David Park was ----- to revise the draft of the contract, making it in the best interest of the company. (A) decided (B) intended (C) offered (D) instructed
A type of the grammar question
To help our staff better understand the nature of the meeting, the agenda was ----- a moment ago. (A) distribute (B) distributing (C) distributed (D) distribution

$$Recall = |A \cap B| / |A|,\ Precision = |A \cap B| / |B| . \quad (3)$$

where, A is the relevant set of sentences(n-gram) for the query, B is the set of retrieved sentences.

$$F1\ measure = 2 \times R \times P / (R + P) . \quad (4)$$

where, R is the Recall rate and P is the Precision rate.

Table 7. Evaluation results

	Measurement	Vocabulary	Grammar	Total
5gram	Recall(%)	56.8	46.667	**53**
	Precision(%)	78.873	100	**85.849**
	F1-measure	66.041	63.636	**65.538**
4gram	Recall(%)	90.16	79.92	**85**
	Precision(%)	85.455	86.667	**85.882**
	F1-measure	87.746	881.504	**85.438**
Trigram	Recall(%)	100	98.611	**99.5**
	Precision(%)	75.781	85.915	**79.397**
	F1-measure	86.222	91.826	**88.318**
Trigram & 4gram	Recall(%)	100	97.436	**99**
	Precision(%)	83.607	86.842	**84.848**
	F1-measure	91.071	91.831	**91.379**

As we can see in table 7, the recall rate based on Google 5gram is only around 53%. In other word, the probability that five continuous words from TOEIC are matched with Google 5gram data is approximately 53%. When we have word set A and B consisted of five continuous words, two word sets are matched by 53%. Although the recall rate is low, the precision rate is nearly 86% which means that if the system found matched 5gram data, this data would close to answer with 86%. The recall rate based on Google 4gram was suddenly increased to around 85% and the precision rate was stayed in steady. The requirement of matching condition using 5gram was five words but 4gram is four words. It is simple to understand that the recall rate based on 4gram was increased to 85%. For the same reason, the recall rate of using trigram was increased to 99.5% which means that most of three continuous words from TOEIC test are placed in Google trigram data sets. However, the precision rate was decreased to 79.397%. It means that there are too many matched trigram data between TOEIC test and Google trigram. Because system based on the logic that it choose the word which has the highest frequency. But the way of natural English language expression is so much dynamic so the word with highest frequency is not always the answer. This is the reason why the precision rate of using trigram is lower than using 4gram and 5gram. It is best if answer word had the highest frequency but is not. To overcome this limitation, we combined 4gram and trigram together due to the fact that 4gram has the highest precision rate with smaller size than 5gram and trigram has the highest recall rate than others. The procedure to find answer has two steps that the system chooses an answer based on 4gram at first. If the system can't find answer, try again using trigram. We believe that this combined method can improve the recall and precision rate both of all. The table 7 supports this point that the performance rates are improved. The first graphs of figure 2 show the recall, precision and F1 rate of vocabulary questions and second graphs give the result of grammar questions. The last one is the final result graphs using trigram and 4gram together to find the answer for TOEIC test.

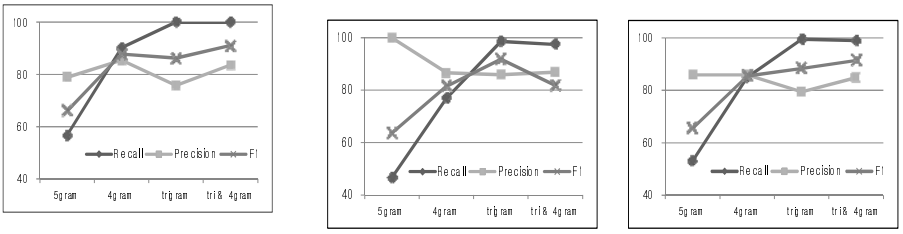

Fig. 2. First Graphs is result using vocabulary, second is grammar questions and third is total result

5 Conclusion and Future works

We evaluated a usability of Google n-gram data through applying n-gram data to part five in TOEIC test sentences provided by Hackers TOEIC. The testing results are based on frequencies that Google has been providing. We have not applied any statistic model to the system yet, but the results of recall and precision rate are reliable when the system using both trigram and 4gram. It has a limitation that the system is not able to find answer if there is no matched n-gram data in Google n-gram. This is means that the candidate n-gram from TOEIC test has to be placed in Google n-gram data to find the answer. To overcome this limitation, we are in progress to apply probability model such as HMM (Hidden Markov Model) and so on. Also, we need to apply our method to another English test data such as TOEFL and IELTS and so on to improve its objectivity and usability. We expect that Google n-gram data has a huge potential that applicable to the query recommendation system, automatic text completion, spell checking and natural language processing and so on. The obstacle for using Google n-gram is the size of data that is too huge. Even we filtered the special character of data, the size of 4gram is nearly 13GB and it is not only unacceptable for real time system but also needs lots of costing time and maintenance efforts. For these reason, the method to reduce size of n-gram is required. We believe that it is possible to reduce the size when we build specified n-gram data to fit user personal interests or characteristics.

Acknowledgments. This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MEST) (No. 2010-0011656).

References

1. Yang, S., Zhu, H., Apostoli, A., Cao, P.: N-gram Statistics in English and Chines: Similarities and Differences. In: International Conference on Semantic Computing, pp. 454–460 (2007)
2. Brown, P.F., de Souza, P.V., Mercer, R.L., Pietra, V.J.D., Lai, J.C.: Class-Based n-gram Models of Natural Language. Computational Linguistics 18(4), 467–479 (1992)
3. Su, Z., Yang, Q., Lu, Y., Zhang, H.: WhatNext: A Prediction System for Web Requests using N-gram Sequence Models. Web Information Systems Engineering 1, 214–221 (2000)

4. Khudanpur, S., Wu, J.: A Maximum Entropy Language Model Integrating n-grams and Topic Dependencies for Conversational Speech Recognition. In: Proceedings of ICASSP 1999, pp. 553–556 (1999)
5. Hwang, M., Choi, D., Choi, J., Lee, H., Kim, P.: Text Editor based on Google Trigram and its Usability. In: UKSim 4th European Modelling Symposium on Computer Modelling and Simulation, pp. 12–15 (2010)
6. Siu, M., Ostendorf, M.: Variable N-Grams and Extensions for Conversational Speech Language Modeling. IEEE Transactions on In Speech and Audio Processing 8(1), 63–75 (2000)
7. Bergsma, S., Lin, D., Goebel, R.: Web-Scale N-gram Models for Lexical Disambiguation. In: Proceedings of the 21st International Joint Conference on Artificial Intelligence, pp. 1507–1512 (2009)
8. Nakov, P., Hearst, M.: Search Engine Statistics Beyond the n-gram: Application to Noun Compound Bracketing. In: Proceedings of the 9th Conference on Computational Natural Language Learning, pp. 17–24 (2005)
9. Siivola, V., Pellom, B.L.: Growing an n-gram language model. In: Proceedings of 9th European Conference on Speech Communication and Technology (2005)