# Measuring Popularity of Machine-Generated Sentences Using Term Count, Document Frequency, and Dependency Language Model

**Jong Myoung Kim**[1], **Hancheol Park**[2], **Young-Seob Jeong**[1]
**Ho-Jin Choi**[1], **Gahgene Gweon**[2], **and Jeong Hur**[3]
[1]School of Computing, KAIST
[2]Department of Knowledge Service Engineering, KAIST
[3]Knowledge Mining Research Team, ETRI
{grayapple, hancheol.park, pinode, hojinc, ggweon}@kaist.ac.kr
jeonghur@etri.re.kr

## Abstract

We investigated the notion of *"popularity"* for machine-generated sentences. We defined a popular sentence as one that contains words that are frequently used, appear in many documents, and contain frequent dependencies. We measured the popularity of sentences based on three components: *content morpheme count, document frequency,* and *dependency relationships.* To consider the characteristics of agglutinative language, we used content morpheme frequency instead of term frequency. The key component in our method is that we use the product of content morpheme count and document frequency to measure word popularity, and apply language models based on dependency relationships to consider popularity from the context of words. We verify that our method accurately reflects popularity by using Pearson correlations. Human evaluation shows that our method has a high correlation with human judgments.

## 1 Introduction

Natural language generation is widely used in variety of Natural Language Processing (NLP) applications. These include paraphrasing, question answering systems, and Machine Translation (MT). To improve the quality of generated sentences, arranging effective evaluation criteria is critical (Callison-Burch et al., 2007).

Numerous previous studies have aimed to evaluate the quality of sentences. The most frequently used evaluation technique is asking judges to score those sentences. Unlike computer algorithms, humans can notice very delicate differences and perceive various characteristics in natural language sentences. Conventional wisdom holds that human judgments represent the gold standard; however, they are prohibitively expensive and time-consuming to obtain.

Because of the high cost of manual evaluation, automatic evaluation techniques are increasingly used. These include very popular techniques that measure meaning adequacy and lexical similarity, such as BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), and TER plus (Snover et al., 2009). Additionally, a distinctive characteristic of auto evaluation techniques is that they can be applied not only to performance verification, but also to the generation stage of NLP applications. Although these techniques can make experiments easier and accelerate progress in a research area, they employ fewer evaluation criteria than humans.

In general, previous research efforts have focused on "technical qualities" such as meaning and grammar. However, customer satisfaction is sometimes determined more by "functional quality" (how the service work was delivered) than by "technical quality" (the quality of the work performed) (Mittal and Lassar, 1998). Especially, Casaló et al. (2008) showed that the customers' loyalty and satisfaction are affected by their past frequent experiences. We focused on this aspect and propose a new criterion, popularity, to consider the functional quality of sentences. We define a popular sentence as one that contains words that are frequently used, appear in many documents, and contain frequent dependencies. Us-



*29th Pacific Asia Conference on Language, Information and Computation: Posters, pages 319 - 327*
*Shanghai, China, October 30 - November 1, 2015*

ing this definition, we aim to measure the popularity of sentences.

In this paper, we investigate the notion of *"popularity"* for machine-generated sentences. We measured popularity of sentences with an automatic method that can be applied to the generation stage of MT or paraphrasing. Because it is a subjective evaluation, measuring the popularity of sentences is a difficult task. We defined a popular sentence as one that contains words that are frequently used, appear in many documents, and contain frequent dependencies. Subsequently, we began our analysis by calculating Term Frequency (TF). To reflect the characteristics of agglutinative languages, we apply a morpheme analysis during language resources generation. As a result, we obtain a Content Morpheme Count (CMC). To complement areas CMC cannot cover (words that have abnormally high CMC), we apply morpheme-based Document Frequency (DF). Lastly, to consider popularity came from contextual information, we apply a dependency relationship language model. We verify our method by analyzing Pearson correlations between human judgments; human evaluation shows that our method has a high correlation with human judgments. And our method shows the potential for measuring popularity by involving the contextual information.

The remainder of this paper is organized as follows. Section 2 presents related works in the field of sentence evaluation. Section 3 explains the approach to measure the popularity of words and sentences. In Section 4, we evaluate the usefulness of our method. In section 5, we analyze the result of experiment Lastly, Section 6 concludes the paper.

## 2 Related Works

Manual evaluation, the most frequently used technique, asks judges to score the quality of sentences. It exhibits effective performance, despite its inherent simplicity. Callison-Burch asked judges to score fluency and adequacy with a 5-point Likert scale (Callison-Burch et al., 2007), and asked judges to score meaning and grammar in a subsequent paper (Callison-Burch, 2008). Similarly, Barzilay et al. asked judges to read hand-crafted and application-crafted paraphrases with corresponding meanings, and to identify which version was most readable and best represented the original meaning (Barzilay and Lee, 2002). Philip M. Mc et al. studied overall quality using four criteria (McCarthy et al., 2009). Using these evaluation techniques, humans can identify characteristics that machines cannot recognize, such as nuances and sarcasm. Overwhelmingly, humans are more sensitive than computers in the area of linguistics. As a result, manual evaluation provides the gold standard. However, manual evaluation presents significant problems. It is prohibitively expensive and time-consuming to obtain.

To address these limitations, there have been studies involving automatic evaluation methods. Papineni et al. (2002) and Callison-Burch et al. (2008) proposed methods that measure meaning adequacy based on an established standard. Several methods based on Levenshtein distance (Levenshtein, 1966) calculate superficial similarity by counting the number of edits required to make two sentences identical (Wagner and Fischer, 1974; Snover et al., 2009). These methods can be used to calculate dissimilarity in paraphrasing. Chen et al. measured paraphrase changes with n-gram (Chen and Dolan, 2011). These automatic evaluations also present a problem — the absence of diversity. There are many senses humans can detect from sentences, even if they are not primary factors such as meaning adequacy or grammar. We identify a novel criteria, popularity, as one of those senses, based on the fact that customer satisfaction is sometimes derived from functional quality (Mittal and Lassar, 1998).

We define the popularity of a sentence using TF, DF and dependency relations. TF, defined as the number of times a term appears, is primarily used to measure a term's significance, especially in information retrieval and text summarization. Since Luhn used total TF as a popularity metric (Luhn, 1957), TF has been frequently used to measure term weight, and employed in various forms to suit specific purposes. Term Frequency-Inversed Document Frequency (TF-IDF), the most well-known variation of TF, is used to identify the most representative term in a document (Salton and Buckley, 1988). Most previous research using those variations has focused on the most significant and impressive terms. There has been minimal research concerned with commonly used terms. We measured popularity of sentences with these commonly used terms that

have high TF and DF.

## 3 Method

In this section, we explain the process of language resource generation, and propose a method to measure the popularity of sentences. First, we utilize morpheme analysis on the corpus of sentences, because our target language is Korean which is an agglutinative languages. Next, we statistically analyze each content morpheme occurrence, and then calculate sentence popularity using these resources.

### 3.1 Korean Morpheme Analysis

We built our language resources (Content Morpheme Count-Document Frequency (CMC-DF) and Dependency Language Model (DLM)) by analyzing a massive corpus of Korean sentences statistically. Because Korean is an agglutinative language, we needed to conduct morpheme analysis before we built those resources. In agglutinative language, words can be divided into content morphemes and an empty morpheme. Content morphemes contain the meaning of words, while empty morphemes are affixed to the content morpheme to determine its grammatical role. For example, in the sentence "뉴욕에 가다. (Go to New York City.)", a word "뉴욕에" can be divided into "뉴욕" and "에". A content morpheme "뉴욕" means "New York city" and an empty morpheme "에" do the role of a stop word "to". Because there are numerous combinations of two morpheme types, it is not appropriate to compile statistics on the words without morpheme analysis. Via this process, we can disassemble a word into a content morpheme and empty morpheme, and obtain a statistical result that accurately represents the word. Postpositions and endings, the stop words of Korean, are filtered in this process. Additionally, we conduct conjunctions filtering, most of stop words of Korean are eliminated in morpheme analysis and filtering. We used a Korean morpheme analyzer module created by the Electronics and Telecommunications Research Institute (ETRI)[1].

### 3.2 Measuring Word Popularity

Before calculating the popularity of sentences, we attempt to measure the popularity of words. We defined a popular word as one with a frequently used content morpheme. The empty morphemes are not considered, because they are stop words in Korean. We adopt Content Morpheme Count (CMC), a variation of TF, to measure usage of the content morpheme of words. CMC is the frequency of a word's content morpheme in a set of documents. The CMC of the word $w$ is driven in the following equations.

$$CMC_w = max(0, \log b(w)) \quad (1)$$

$$b(w) = \sum_{d \in D} f_{m,d} \quad (2)$$

In Eq. (2), $b(w)$ is the qualified popularity of word $w$, defined as the number of content morphemes $m$ of word $w$ in entire documents $D$. $f$ is the frequency of a particular content morpheme $m$ in document $d$. We applied the logarithm in Eq. (1) because simple frequency measures have a tendency to emphasize high-frequency terms. Furthermore, we utilize the max function to handle unseen morphemes in the training data corpus.

$$B(s) = \frac{\sum_{i=1}^{n} CMC_{w_i}}{n} \quad (3)$$

Using the average of CMC, we measure the popularity $B(s)$ of sentences with a size of $n$ in Eq. (3). We use this score as a baseline.

Unfortunately, CMC is not sufficient to reflect the popularity of words, because there are some cases it cannot cover. Technical terms or named entities frequently occur only in a few documents such as scientific articles or encyclopedia entries. In those cases, a high CMC is calculated, even if those words are not popular to the general public. For example, the word "스타매거진 (star magazine)" occurred 270 times in only one document (Korean Wikipedia contains 700,000 documents). Similarly, the word "글루코코르티코이드 (glucocorticoid)" occurred 39 times in only one document (the common word "카펫 (carpet)" occurred 41 times over 36 documents). The CMCs of those words are relatively high, but they are not popular terms to ordinary people.

Thus, we inversely applied the concept of TF-IDF. TF-IDF assumes that if a term frequently occurs in only a few documents, the term is significant in those documents. Inversely, we assumed that if a content

[1] https://www.etri.re.kr/kor/main/main.etri

morpheme is used frequently and occurs in numerous documents, that morpheme is popular to people. We quantified the popularity of words as the number of documents in which its content morpheme occurs. We calculate Document Frequency (DF) in the following equations.

$$DF_w = max(0, \log c(w)) \tag{4}$$

$$c(w) = \mid d \in D : m \in d \mid \tag{5}$$

In Eq. (5), $c(w)$ is the number of documents $d$ in which content morpheme $m$ occurs. Similarly, logarithm and max function are applied in Eq. (4).

Using the notion of CMC and DF, we defined the popularity of words $f(w)$ as follows in Eq. (6). Lastly, we measured the popularity of sentences by calculating the average popularity of words in sentences with a size of $n$. Popularity of sentences $F(s)$ based on word popularity is represented in Eq. (7)

$$f(w) = CMC_w \times DF_w \tag{6}$$

$$F(s) = \frac{\sum_{i=1}^{n} f(w_i)}{n} \tag{7}$$

### 3.3 Measuring Context Popularity

We measure popularity of sentence using word popularity in the previous section. Nevertheless, there is another element as important as word popularity. The element is whether the word is suitable in the context. For example, we frequently use "powerful," not "strong," when discussing a computer having substantial computational ability. As an adjective, "powerful" and "strong" have similar meanings and popularity. The use of "strong" is perhaps more frequent than that of "powerful." However, if we consider the context of a noun "computer" joined with each word, i.e., "powerful computer" and "strong computer", there will be a significant difference in popularity of the two phrases.

To address this aspect, we observe a word that has a direct semantic relationship with target word. The word is dependency head of target word. In the sentence, every word (except the dependency root word) has a dependency head, and it is related to the head. We attempt to verify the potential that context can influence popularity of sentence with dependency head. We created a Dependency Language Model (DLM), which reflects the conditional probability of words and its dependency head.

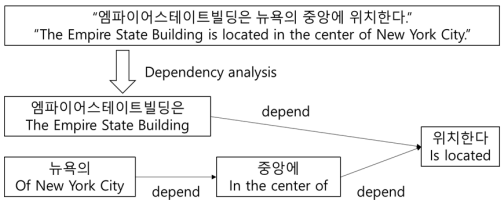


Figure 1: Example of dependency analysis in a Korean sentence.

We obtained the probability from a frequency investigation of a word pair after a dependency analysis. For example, the sentence "엠파이어스테이트빌딩은 뉴욕의 중앙에 위치한다. (The Empire State Building is located in the center of New York City.)" can be disassembled into {"엠파이어스테이트빌딩은 (The Empire State building)" → "위치한다 (is located)"}, {"뉴욕의 (of New York City)"→"중앙에 (in the center)"} and {"중앙에 (in the center)"→"위치한다 (is located)"}. This process is represented in Figure 1. Then, we investigated the conditional probabilities of those pairs. Thus, we calculate the conditional probability of words pairs as a unit of DLM. In addition, we applied morpheme analysis for the reasons described in Section 3.1. We used the dependency analyzer created by ETRI.

$$p(w|h_w) = \frac{CMC_{w,h_w}}{CMC_{h_w}} \tag{8}$$

Eq. (8) represents the conditional probability $p(w|h_w)$ of word $w$ and its head $h_w$. $CMC_{w,h_w}$ is the number of co-occurrence of $w$ and $h_w$. DLM is built by investigating all the dependency pairs of the corpus. Using the notion of DLM, we defined the context popularity $g(w)$ as product of two words popularity (target word and its head) and their co-occurrence probability. It is represented in Eq. (9).

$$g(w) = f(w)p(w|h_w)f(h_w) \tag{9}$$

To measure sentence popularity with DLM, we calculate the context popularity of all dependency word pairs. This process is represented by the formula in Eq. (10). Lastly, to normalize the length $n$

of sentences, we apply a logarithm and divide it by the number of dependency relationships $n-1$.

$$D(s) = \prod_{i}^{n-1} f(w_i)p(w|h_{w_i})f(h_{w_i}) \quad (10)$$

$$G(s) = \frac{\sum_{i}^{n-1} \log g(w)}{n-1} \quad (11)$$

Additionally, we can treat the word sense disambiguation (WSD) problem, too. For example, in the Korean language, the meaning of the noun "배 [bæ]" may be "pear" or "boat." When we analyze the corpus to build CMC and DF, both nouns are treated as a single entity. This results in abnormally high statistical result scores, regardless of the actual frequency of each meaning. Using DLM, we can consider this problem with conditional probability. For example, the noun "배" means "pear" when its dependency head is *eat* or *squash*, and means "boat" if it is matched with *sail* or *steer*. We can infer the meaning and popularity of words in the context from its dependency head.

### 3.4 Measuring Total Popularity

We defined a popular sentence as one that contains words that are frequently used, appear in many documents, and contain frequent dependencies. In Eq. (12), we represent the sentence popularity $H(s)$ by the sum of popularities from words $F(s)$ and popularities from contexts $G(s)$. In the equation, $\alpha$ and $\beta$ are the weights of both popularities.

$$H(s) = \alpha F(s) + \beta G(s) \quad (12)$$

We can obtain Eq. (13) through substitution of Eq. (7) and (11) into Eq. (12).

$$H(s) = \alpha \frac{\sum_{i=1}^{n} f(w_i)}{n} + \beta \frac{\sum_{i}^{n-1} \log g(w_i)}{n-1} \quad (13)$$

## 4 Experimental Setup

To evaluate how accurately our metric reflects the popularity that humans perceive when reading a sentence, we designed an experiment to measure correlation between human judgment and popularity.

### 4.1 Adoption of Dataset

To build the CMC, DF, and DLM, we need an appropriate corpus. When searching for a target corpus, the most important considerations were volume and ordinariness. Thus, we considered *Korean Wikipedia*[2] and *Modern Korean Usage Frequency Report (MKUFR)* (Hansaem, 2005) as suitable data sources. Korean Wikipedia is the Korean version of Wikipedia, the well-known collaborative online encyclopedia. Because its is written by public, we assume Korean Wikipedia contains moderately popular terms. Korean Wikipedia even offers a massive volume — more than 1.7 GB of data contained in over 700,000 documents. MKUFR is the result of research conducted by the National Institute of the Korean language from 2002 through 2005. They surveyed TF in publications printed between 1990 and 2002. Using Korean Wikipedia and MKUFR as a dataset, we built the CMC, DF, and DLM.

### 4.2 Human Evaluation Setup

We used a sentence set from TREC 2006 QA data[3] as test data. TREC (Text REtrieval Conference) is a conference focusing on information retrieval areas, and its dataset is widely used as a standard to evaluate the performance of information retrieval systems. We randomly selected 250 sentences from the TREC 2006 QA data and translated them into Korean by human translators. A paraphrase machine, based on Bannard and Callison-Burch's algorithm (Bannard and Callison-Burch, 2005), was used to create machine-generated sentences from the translated TREC questions.

We employed five human judges (J1-5) to manually assess the popularity of 250 machine-generated sentences. The sentences were presented to the judges in random order. Each sentence was scored using a six-point scale. The instructions given to the judges were as follows.

*Popularity: Is the sentences linguistically popular?*

### 4.3 Inter-judge Correlation

Before evaluating our method, we used Pearson's correlation coefficient to investigate the correlation between the human judges; these results are listed

[2]https://ko.wikipedia.org

[3]http://trec.nist.gov/data/qa/

| | J1 | J2 | J3 | J4 | J5 |
|---|---|---|---|---|---|
| J1 | 1 | 0.639 | 0.722 | 0.650 | 0.639 |
| J2 | 0.639 | 1 | 0.582 | 0.496 | 0.645 |
| J3 | 0.722 | 0.582 | 1 | 0.724 | 0.638 |
| J4 | 0.650 | 0.496 | 0.724 | 1 | 0.536 |
| J5 | 0.639 | 0.645 | 0.638 | 0.536 | 1 |

Table 1: Inter-judge correlation.

| | | J avg. | J1 | J2 | J3 | J4 | J5 |
|---|---|---|---|---|---|---|---|
| Wiki | CMC | .45 | .40 | .37 | .39 | .33 | .39 |
| | CMCDF | **.58** | **.53** | **.51** | **.50** | **.40** | **.50** |
| UFR | CMC | .30 | .28 | .28 | .24 | .19 | .27 |
| | CMCDF | .43 | .37 | .40 | .38 | .27 | .37 |

Table 2: Correlation between human judgment and popularity of each corpus.

in Table 1. Although J4 produced relatively poor results, correlations show a clear positive relationship between 0.49 and 0.72; excepting J4's results, the correlation improved to between 0.58 and 0.72. These correlation scores can be regarded as fairly high, considering that we used a six-point scale and compared the results to similar results reported during the paraphrase evaluation (Liu et al., 2010). These high correlations confirm the effectiveness of our experimental design and explanation. We considered the reasons for J4's relatively poor score when analyzing the results.

## 5 Experimental Result

### 5.1 Word Popularity

To measure sentence popularity with word popularity, we built language resources (CMC and DF) from each corpus: *Korean Wikipedia* and *Modern Korean usage frequency report*. Using Eq. (3) and Eq. (7), we calculated the popularity of each sentence. By comparing the performance of each corpus, we aim to identify the corpus that most accurately reflects public language usage. The correlations between our method and human judgments are listed in Table 2.

The results in Table 2 show, in general, clear positive linear correlations. The row labeled "Wiki" shows the results based on the Korean Wikipedia corpus; row "UFR" shows results based on MKUFR. In particular, Wikipedia rather than MKUFR shows better performance, and CMC-DF (represented in Eq. (7)) shows better performance than CMC (Eq. (3)) only. We conclude that Korean Wikipedia reflects public language usage more accurately than MKUFR. Thus, we selected the Wikipedia corpus as the basis of our DLM, and conducted the experiment described below.

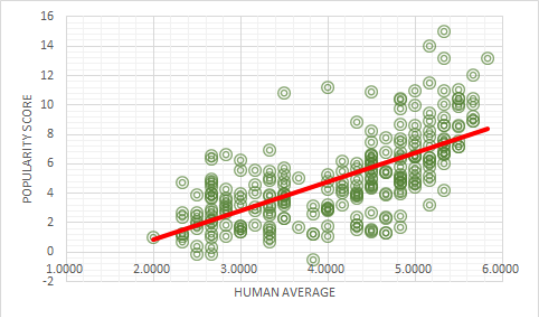


Figure 2: Scatter plot of popularity (un-optimized) versus human judgment (avg.).

### 5.2 Context Popularity

Through the previous experiment, we conclude that the *Wikipedia corpus* most accurately reflects public language usage. Thus, we built a DLM based on Wikipedia. Using Eq. (11), we measured context popularity based on dependency relationships. Lastly, we attempted to measure popularity by applying both word popularity and context popularity (this process is represented in Eq. (13)). In the Table 3, row "DLM" contains the results of applying context popularity (represented in Eq. (11)); row "Comb" contains the results of applying both word popularity and context popularity (represented in Eq. (13)). Figure 2 shows the average of human judgment scores plotted against the popularity derived from Eq. (13). Lastly, the results in row "Opt" show the result of optimization of the weight variables $\alpha$ and $\beta$ of Eq. (13). The optimization process will be discussed in Section 5.3. An interesting finding is that considering contexts alone is negatively correlated with human judgments. Nevertheless, when they are combined with word popularity, performance is improved. The Pearson correlation between popularity and human judgment is 0.77.

| | Javg. | J1 | J2 | J3 | J4 | J5 |
|---|---|---|---|---|---|---|
| CMC | .45 | .40 | .37 | .39 | .33 | .39 |
| CMCDF | .58 | .53 | .51 | .50 | .40 | .50 |
| DLM | -.20 | -.23 | -.17 | -.15 | -.17 | -.22 |
| Comb | .66 | .62 | .60 | .56 | .44 | .62 |
| **Opt** | **.77** | **.69** | **.60** | **.67** | **.45** | **.80** |

Table 3: Correlation between human judgment and popularity of different models.

### 5.3 Weight Optimization

To derive optimal weight parameter $\alpha$ and $\beta$ in Eq. (13), we divide the experiment data into three sets: training, validation, and test. We divided the experiment data using a ratio of $3 : 1 : 1$. Using a grid search, we identify the top ten parameter combinations. We set the scope of each parameter as integer $[0, 100]$. By applying those combinations to the validation set, we identify the optimal parameter pair; the parameter of CMC-DF($\alpha$) is 35 and that of DLM($\beta$) is 65. We verified the performance of our method with test set using the optimal parameter pair obtained from the validation set ($\alpha : \beta = 35 : 65$).

### 5.4 Result Analysis

As in the Section 5.1, we investigated CMC-DF's Pearson correlation with human judgments. Our basic concept started with term frequency; we built language resources (CMC) based on term frequency, and they showed a clear positive correlation of 0.45. In addition, we suggested that cases cannot be solved using only CMC. Thus, we applied DF, and we obtained an improved correlation of 0.58.

To measure popularity stemming from contextual information, we applied language modeling based on dependency relationships. Interestingly, DLM shows negative correlation by itself. However, when combined with CMC-DF, it improves correlation; the Pearson correlation between the combined model (CMC-DF-DLM) and human judgment is 0.66. We optimized the weight parameters through a grid search and avoid overfitting by dividing experiment data into three categories: training, validation, and test. The Pearson correlation between our popularity method and human judgment is 0.77. This correlation is quite high, considering that the highest sentence-level Pearson correlation in the MetricMATR 2008 (Przybocki et al., 2009) competition was 0.68, which was achieved by METEOR; in contrast, BLEW showed a correlation of 0.45. When compared with the results of PEM (Liu et al., 2010), the sentence level correlation is also quite high.

Furthermore, we calculated the correlation between our method and each judge. Except for one judge, our method shows strong positive linear correlation with human judgments (between 0.60 and 0.80). Although the results produced by J4 were relatively poor, they still resulted in a clear positive correlation of 0.45.

### 5.5 Characteristics in Corpora and Judges

Table 2 shows that the CMC-DF based on Korean Wikipedia exhibit better performance than those based on MKUFR. The results from Section 5.1 became our grounds for concluding that Korean Wikipedia reflects public language usage more accurately than MKUFR. We believe the reasons are as follows.

- *Wikipedia* is written by the public.
  Modern Korean usage frequency report is based on publications written by experts such as writers, journalists, novelists, etc.

- *Wikipedia* is written in real time.
  Modern Korean usage frequency report was created in 2005 and analyzed publications printed between 1990 and 2002.

In Table 1, 2 and 3, we note that J4's results show relatively low correlation with the results from other judges and the results from our methods. To reveal the reason, we analyzed their answer sheets. Table 4 shows statistical characteristic of human judgments. For each judge's decision, $\mu$ is the average score, $\sigma$ represents the standard deviation, and *min* and *max* represent the lowest and highest values, respectively, in the range of responses. The salient point in Table 4 is that J4 assigned scores in a range of only $[2, 5]$ while others used the entire scale $[1, 6]$.

### 5.6 Discussion and Future Work

As shown in Table 2 and 3, our method shows strong correlations with human judgments, even in

| | Javg. | J1 | J2 | J3 | J4 | J5 |
|---|---|---|---|---|---|---|
| $\mu$ | 4.11 | 3.35 | 4.69 | 3.74 | 4.41 | 3.02 |
| $\sigma$ | 0.99 | 1.21 | 1.05 | 1.86 | 0.91 | 1.31 |
| *min* | 2.00 | 1 | 1 | 1 | **2** | 1 |
| *max* | 5.83 | 6 | 6 | 6 | **5** | 6 |

Table 4: Comparison of statistical characteristics of human judgments.

cases in which differences exist between individuals. Further, there is a clear improvement in correlation when the additional notions of document frequency and context are applied. In this experiment, our method showed the potential for measuring popularity by involving the contextual information; so far, we have considered only one word that has direct semantic relationship with target word, namely, the dependency head. The extension of contextual information will be addressed in future works.

Our method has a limitation due to lexical features. We cannot accommodate syntax-level popularity measures, such as the order of words. Because Korean is affiliated with agglutinative languages, there is no grammatical or semantic meaning related to the order of words; in sentences, empty morphemes decide the role of content morphemes. However, for readers and service consumers, the order of words can convey different impressions. This is an extension of the characteristics we aim to measure using popularity. These types of syntactic factors will be addressed in future works.

J4 showed relatively low Pearson correlation performance, breadth of improvement, and inter-judge correlation. To explain these, we developed two hypotheses. The first is that J4 assigned scores in a range of only $[2, 5]$ while others used the entire scale $[1, 6]$. When conducting an experiment using the Likert scale, it is common for judges to avoid extreme estimations. This can reduce the sensitivity of the results. Low inter-judge correlation supports this hypothesis. The second is that he had a different standard of popularity. As mentioned previously, popularity is very subjective sense, and we focused on popularity stemming from lexical factors. If he followed different rules than other judges, the relatively low performance can be explained. Low improvement breadth per application of additional factors supports this hypothesis.

In aspects of application, we consider popularity as a method to reflect the style of sentences produced by MT or paraphrasing methods. Popularity is a type of combination of weighted probabilities. This means generating a possibility under a corpus that accurately reflects a target. In this paper, the target of the corpus was public language usage. However, if we secure various corpora that each reflects different targets, they can be used as classifiers to find the author of the source sentences.

Further, resources (CMC, DF, and DLM) can be used for generation module of MT or paraphrase system to reflect the specificity of the author. The target of corpus can be time, author, topic, or other factors. For example, assume that we have obtained diverse corpora from various authors, and one of the authors, "Murakami Haruki," writes a new novel. A MT system containing the popularity module and language resources can identify the novel's author and apply his style using the language resources from the Murakami's corpus in generation stage.

## 6 Conclusion

In this paper, we proposed a novel notion, popularity, to consider the consumers' satisfaction from functional quality. We defined a popular sentence as one that contains words that are frequently used, appear in many documents, and contain frequent dependencies. To measure the popularity, we began with term frequency, and then applied the concepts of document frequency and context to complement features that term frequency cannot cover. We conducted a human evaluation and measured the popularity for machine-generated sentences.In our experiment, we showed strong Pearson correlation coefficients between popularity and human judgment. To the best of our knowledge, our method is the first automatic sentence popularity evaluator based on term occurrences and contextual information.

## ACKNOWLEDGMENTS

This work was supported by ICT R&D program of MSIP/IITP. [R0101-15-0062, Development of Knowledge Evolutionary WiseQA Platform Technology for Human Knowledge Augmented Services]

## References


Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, volume 29, pages 65–72.

Colin Bannard and Chris Callison-Burch. 2005. Paraphrasing with bilingual parallel corpora. In *Proceedings of the 43rd Annual Meeting on ACL*, pages 597–604.

Regina Barzilay and Lillian Lee. 2002. Bootstrapping lexical choice via multiple-sequence alignment. In *Proceedings of the ACL-02 conference on EMNLP-Volume 10*, pages 164–171.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2007. (meta-) evaluation of machine translation. In *Proceedings of the Second Workshop on StatMT*, pages 136–158.

Chris Callison-Burch, Trevor Cohn, and Mirella Lapata. 2008. Parametric: An automatic evaluation metric for paraphrasing. In *Proceedings of the 22nd Coling*, pages 97–104.

Chris Callison-Burch. 2008. Syntactic constraints on paraphrases extracted from parallel corpora. In *Proceedings of the Conference on EMNLP*, pages 196–205.

Luis Casaló, Carlos Flavián, and Miguel Guinalíu. 2008. The role of perceived usability, reputation, satisfaction and consumer familiarity on the website loyalty formation process. *Computers in Human Behavior*, 24(2):325–345.

David L Chen and William B Dolan. 2011. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th Annual Meeting on the ACL*, pages 190–200.

Kim Hansaem. 2005. Modern korean usage frequency report. `https://books.google.co.kr/books?id=umhKAQAAIAAJ`.

Vladimir I Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet Physics Doklady*, volume 10, pages 707–710.

Chang Liu, Daniel Dahlmeier, and Hwee Tou Ng. 2010. Pem: A paraphrase evaluation metric exploiting parallel texts. In *Proceedings of EMNLP*, pages 923–932.

Hans Peter Luhn. 1957. A statistical approach to mechanized encoding and searching of literary information. *IBM Journal of Research and Development*, 1(4):309–317.

Philip M McCarthy, Rebekah H Guess, and Danielle S McNamara. 2009. The components of paraphrase evaluations. *Behavior Research Methods*, 41(3):682–690.

Banwari Mittal and Walfried M Lassar. 1998. Why do customers switch? the dynamics of satisfaction versus loyalty. *Journal of Services Marketing*, 12(3):177–194.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on ACL*, pages 311–318.

Mark Przybocki, Kay Peterson, Sébastien Bronsart, and Gregory Sanders. 2009. The nist 2008 metrics for machine translation challenge—overview, methodology, metrics, and results. *Machine Translation*, 23(2-3):71–103.

Gerard Salton and Christopher Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5):513–523.

Matthew G Snover, Nitin Madnani, Bonnie Dorr, and Richard Schwartz. 2009. Ter-plus: paraphrase, semantic, and alignment enhancements to translation edit rate. *Machine Translation*, 23(2-3):117–127.

Robert A Wagner and Michael J Fischer. 1974. The string-to-string correction problem. *Journal of the ACM (JACM)*, 21(1):168–173.