Idiom Recommendation System

2022310427 Lee Chae Eun
Applied Artificial Intelligence

2022312957 Cho Jun Beom
Applied Artificial Intelligence

1. Introduction

1.1. Topic and Problem

In modern society, effective and concise communication across various environments is essential. Particularly when conveying complex concepts or explaining subtle situations, simple words or sentences often prove insufficient. In such contexts, idioms based on Chinese characters, which often called as '고사성어' in Korean, can convey meanings beyond mere words, allowing for concise yet richly expressive descriptions that leave a profound impression on the listener. However, finding an appropriate idiom for a specific situation is not easy for individuals. Unlike general word searches, idioms require a precise understanding of their meanings and usage contexts to be used correctly. If one attempts to utilize idioms in tasks such as document writing, report preparation, speech crafting, or developing marketing messages with only a vague grasp of their meanings, the message may become distorted and cause confusion. This not only involves the hassle of having to manually search for suitable idioms but also poses the challenge of determining whether the selected expressions are appropriate. To address this issue, this project has chosen the theme of an 'Idiom Recommendation System' with the aim of implementing a system that recommends appropriate classical Chinese idioms for various situations.

1.2. Purpose and Necessity

The main object of this project is to support users in easily finding appropriate idioms that suit their situations. This can provide substantial benefits to both individuals and businesses.

Firstly, it can contribute to enhancing communication skills. Our system can significantly suitable expressions that can describe complex situations more concisely and effectively. By appropriately utilizing idioms in tasks such as document writing, speech preparation, and report creation, the delivery of messages can be strengthened. For example, an expression like '누란지위' (a situation of severe danger) can concisely explain serious risks and leave a strong impression on the audience. By conveniently recommending such appropriate idioms, users can save time and effort while enhancing the quality of their documents and their communication skills.

Secondly, it has high applicability in business and marketing. Idioms can play an important role not only as linguistic expressions but also in business and marketing strategies. Particularly in global marketing, utilizing idioms to present slogans or advertising phrases can add depth to messages and emphasize cultural uniqueness. For instance, appropriately using idioms in exchanges between Korea, China, and Japan can strengthen cultural connections between these countries and contribute to the promotion of cultural exchange. Additionally, employing idioms in brand messages or slogans can leave a more powerful impression on consumers. This approach can strengthen brand identity and enhance competitiveness in the global market.

1.3. Overview of the Service

The flow of our Idiom Recommendation System is composed of detailed stages, ranging from user input to the recommendation of appropriate idioms.

1) Input Data Analysis

The system analyzes the text data provided by the user to extract key words and meanings. By utilizing morphological analysis tools, it identifies major vocabulary such as nouns, verbs, and

adjectives to understand the topic of the input text. This serves as the foundational step for recommending suitable idioms for each situation.

2) Application of Recommendation Algorithms

Based on the extracted keywords, the system applies word embedding techniques to calculate the semantic similarity between the input and pre-saved meanings. This enables the effective recommendation of idioms that can be used in context similar to the input text.

3) Provision of User-Tailored Results

Based on the analysis results, the system offers the user several of the most appropriate idioms. This helps users assess the suitability of the recommended expressions and apply them effectively in real-life situations.

2. Data

2.1. Data Acquisition

To collect data for the project, 2,010 idioms and each of their detailed information was collected from Naver Hanja (Chinese Character) Dictionary (https://hanja.dict.naver.com/#/category/subject). Since our focus is solely on the meaning for training and analysis, the components of each idiom were separated in advance. Additionally, to provide users with supplementary information when needed, the source URL for each idiom was included. This inclusion allows users to access the original context and background of each idiom, facilitating a deeper understanding and more accurate usage. In cases where an idiom has multiple meanings, all meanings were stored together but distinguished in a way that allows both meanings to be considered in subsequent analyses.

2.2. Data description

As previously explained, since the analysis and training will be meaning-centric, we extracted each components that composes each idiom separately. Later we intend to integrate them again using their respective index numbers. The table following is an example of segmented data for idiom '고진감래'.

Components	Content
Hanja	苦盡甘來
Korean	고진감래
Meaning	'쓴 것이 다하면 단 것이 온다.'는 뜻으로, 고생 끝에 즐거움이 옴을 이르는 말.
URL	entry/ccko/11802e80fa60469ebed259a871c48c50

2.3. Data Preprocessing

When reviewing 2,010 idioms provided by the dictionary, we found that some words were duplicated. To address this, we iterated through all the lists to identify and remove duplicate indices of each component lists. This process ensures that the system does not repeatedly present the same word when recommending idioms in the future, thereby offering more diverse and varied information. As a result, by removing 11 duplicated words, we were able to obtain 1,999 idioms stably.

The collected meaning information, which is in a form of sentence, is written in various styles and formats to ensure that users of the dictionary service can understand it quickly and easily. This means it does not follow a consistent structure or set of rules. For instance, even when conveying the same meaning, the sentence structure may differ, or a variety of vocabulary might be used. While such unstructured and flexible expression methods are familiar and easily understood by human readers, it poses challenges for computers to consistently process this diverse data. Therefore, through a preprocessing phase, we aim to eliminate unnecessary parts and extract only the content relevant for subsequent part-of-speech analysis, such as main meanings and origins.

This stage consists of several detailed steps. Firstly, stopwords are removed to filter out unnecessary words for analysis. By eleminating such as '또는' (or), '즉' (that is), '흔히' (commonly), '줄여서' (abbreviated as), and '따위' (etc.), the accuracy of semantic analysis is enhanced, allowing focus to be placed on key terms.

Second, sentences are categorized based on whether they include the phrase '뜻으로' (as the

meaning of). In cases where ‘뜻으로’ is present, the core meaning is contained in the portion before of that word, and special symbols such as quotation marks are removed to effectively extract this meaning. Additionally, since the portion following ‘뜻으로’ may contain supplementary information, phrases that indicate the origin, like ‘에 나오는’ (appearing in)’, ‘에서 나온’ (originating from), and ‘유래’ (origin), as well as words that suggest additional explanations, such as ‘이르’, ‘이른’ (to mean)’, ‘말’, ‘뜻함’ (means) are used to find additional elements for detailed implementation.

Also for sentences that do not include ‘뜻으로’, the sentences are separated based on periods, and important parts are extracted and stored using the same process as described above. This structured approach ensures that most of the relevant meanings and origins are retained, facilitating accurate part-of-speech in subsequent stages. The table below compares the results before and after the process using the meaning of ‘고진감래’ as an example.

Before	‘쓴 것이 다하면 단 것이 온다.’는 뜻으로, 고생 끝에 즐거움이 옴을 이르는 말.
After	‘쓴 것이 다하면 단 것이 온다.’, '고생 끝에 즐거움이 옴을'

3. Methodology and Toolkits

3.1. Morpheme Analysis

As previously mentioned, short sentences containing the core meanings and supplementary elements were extracted from the entire text. In this stage, morphological analysis is performed to extract only the significant word-level components from each sentence, aiming to reduce the size of the data and enhance the accuracy of the training. Especially since the project involves processing Korean text data, we intend to utilize the OKT (Open Korean Text) library. OKT is a morphological analyzer specifically designed for the Korean language, which will assist in accurately extracting and analyzing the essential linguistic components from our data.

The process aims to extract words which has particular POS (part-of-speech), enabling the removal of irrelevant terms and focusing on those that contribute significantly to the analysis. Specifically, verbs, nouns, and adjectives are selected because they are the primary carriers of semantic meaning in sentences. Nouns represent entities and subjects, verbs denote actions and processes, and adjectives describe qualities and attributes. By concentrating on these key parts of POS, the analysis can effectively capture the essential elements of the text, facilitating a deeper understanding of the underlying meanings and relationships.

Additionally, to prevent certain POS from being incorrectly classified as others and included in the results, thereby adversely affecting the analysis, an extra stopword removal step was also implemented. The table below illustrates the results of extracting only significant words that contains specific POS.

Before	‘쓴 것이 다하면 단 것이 온다.’, '고생 끝에 즐거움이 옴을'
After	'쓴 하면 단 온다', '고생 끝 즐거움 옴'

3.2. Word Embedding

After the stage of morpheme analysis, the next crucial step involves implementing word embedding. Word embedding is essential because it transform these linguistic components into dense numerical vectors that encapsulate their semantic meanings and contextual relationships. This numerical representation allows the system to comprehend the nuanced meanings and associations between different words, enabling more accurate and meaningful analysis. In the context of our system, word embedding facilitates the calculation of semantic similarities between the input and existing idioms.

Among various word embedding techniques, we chose to utilize FastText due to its advanced capabilities in handling morphological variation and OOV (out-of-vocabulary) words. Comparing to Word2Vec, the latter treats each words as an indivisible atomic unit, meaning that any word not present in the training corpus lacks a corresponding vector representation. In contrast, FastText employs

a subword approach by breaking down words into smaller units called n-grams. This method allows it to generate meaningful vector representations for OOV words by leveraging the information from their constituents.

In our FastText model, we set the vector size to 300, which provides a rich and detailed representation of each word's semantic meaning. The window size is configured to 5, allowing the model to consider five words to both sides. This broader context helps in understanding the surrounding words and enhances the model's ability to capture meaningful associations. We chose a minimum count of 1 to ensure that all words, regardless of their frequency, are included in the vocabulary as we already extracted important key words from the previous process. The training algorithm is set to Skip-gram, which is effective for leaning high quality embeddings, for infrequent or rare words.

While each word's vector size is consistently set to 300, the number of vectors per meaning sentence varies. This inconsistency makes distance calculations and comparisons challenging when new inputs are introduced. Such an imbalance hinders the consistent measurement of similarity between sentences. To address this problem, we averaged the multiple vectors present in each sentence (or list), consolidating them into a single fixed 300 dimension vector. By doing so, all sentences are represented by vectors of the same size, enabling effective distance calculations and similarity comparisons between vectors. Additionally, the average vector reflects the comprehensive meanings of the words within a sentence, allowing for more accurate and consistent analysis. Through this approach, our service can promptly and accurately recommend similar sentences even for new inputs, thereby enhancing the overall quality and efficiency of the recommendations.

3.3. Idiom Recommendation

Each word was converted into a vector and average to represent each meaning sentence as a single vector. Therefore, when a new search term or sentence is entered, it is numerically represented using the previously trained model. By calculating the distance between this new vector and each existing meaning vector, the system can identify and return the closest matching idiom. Moreover, as we adopted to use FastText algorithm to compose the model, this provides accurate and contextually relevant suggestions, even for inputs that were not encountered during the training phase.

4. Analysis Results

To demonstrate the effectiveness of this service, we plan to select topics that are commonly used in everyday conversations or sentence constructions. These selected topics will be inputted into the system, and we will observe the results it returns. We have designed to return the top 5 idioms because it can focus on delivering the most semantically similar and contextually appropriate idioms, while still offering a variety that caters to different nuances within the user's query. The table below contains English-translated meanings, actually the system returns it in Korean and Chinese letters.

Input : 친구, 우정 (Friend, Friendship)	
金蘭之契	very thick bond between friends.
金蘭之交	
知己	a friend who really understand him/her
知己之友	
傾蓋如舊	first time to met, being close as old friends

Input : 은혜에 보답하고 싶어 (Want to return for favor)	
結草報恩	paying back grace after death
反哺報恩	grows up repay the parent's grace
罔極之恩	limitless grace by the king or parent
寸草心	heart of child to return the favor
刻骨難忘	the great grace that is engraved in the bones

Input : 너무 바빠서 정신이 없어 (I'm so busy)	
有備無患	nothing to worry if prepared
唯一無二	there is only one
鰥寡孤獨	lonely and unreliable situation
水魚之交	an inseparable relationship
無所不知	there's nothing you don't know

By analyzing how the system responds to these frequently encountered subjects, we aim to assess its accuracy, relevance, and overall performance in providing appropriate idiom recommendations. As a result, we found out that our service meets most of

the needs of users in real-word scenarios when clear key words were given. However if sentences which are ambiguous or figurative (정신이 없어), or contains words that are rare or hard to understand are given, idioms that are not very relevant can be sometimes returned, leading to inadequate result.

5. Expected Outcomes

This project offers numerous advantages that facilitate smooth and effective communication. By utilizing idioms which is recommended for its situation, it enables users to convey their complex concepts and emotions in a concise and clear manner, reinforcing the clarity of messages and reducing the potential for misunderstandings.

From a social and cultural standpoint, the service holds the potential for widespread application across various fields. For instance, in the realm of corporate marketing, using idioms in slogans or advertising messages can leave a lasting impression on customers and strengthen the brand's image and identity. In the educational sector, idioms can be used to bridge traditional wisdom with modern contexts, aiding students in learning historical and cultural heritages and enabling them to apply this knowledge creatively. Additionally, our service can contribute to the creation of popular content and cultural content development, supporting innovative communication methods on social media platforms and digital channels.

Notably, idioms can serve as cultural commonalities among East Asian countries such as Korea, China, and Japan. Thus, this service can function not only as a communication tool but also as a medium for cultural exchange and collaboration. For example, projects based on shared idioms within the East Asian cultural sphere can promote mutual understanding and cooperation between these nations. Furthermore, the system can be applied to the development of global marketing messages or utilized as a tool to introduce cultures to audience outside the region.

6. Limitations

Firstly, although this service successfully eliminated duplicate entries and enhanced data consistency through preprocessing during the data collection process, there remains a possibility that it does not fully encompass the diverse contextual uses of idioms. Specifically, relying solely on data collected from Naver Hanja Dictionary presents limitations in providing idioms suitable for all possible situations. This constraint can lead to discrepancies between the contexts entered by users and the idioms recommended by the system.

Second, despite leveraging morphological analysis and word embedding to comprehend the meanings of the text, there are still inherent limitations in perfectly interpreting the complexities and subtle contextual variations of natural language. Particularly, the system may misinterpret polysemous expressions, metaphorical phrases, and sentences containing conflicting or unrelated words, resulting in the recommendation of inappropriate idioms. To overcome these limitations, it is necessary to incorporate additional embedding techniques, such as Sentence-BERT, which can capture meaning at the sentence level.

The FastText model provides the capability to generate embeddings based on subword information for OOV words. However, lastly, for words that are either insufficiently observed in the training data or are not frequently used in specific contexts, the accuracy of the generated embeddings may decrease. To mitigate these issues, it demands a larger FastText model, which was pre-trained by others, or to collect more extensive datasets that encompass a wide variety of words and contexts. This approach will improve the quality of word embeddings and enhance the model's generalization performance.

7. Conclusion

As highlighted in the introduction, our idiom recommendation system offers a solution for increasingly important effective communication in modern society by utilizing idioms. This system is particularly adept at conveying complex concepts

and explaining subtle situations in a concise yet impactful manner. It can significantly enhance communication across various domains, ranging from interpersonal interactions to corporate marketing messages, and further contribute to improving the efficiency of societal communication. Moreover it can bridge groups connected by culture and history, serving as a common denominator for fostering mutual understanding and consensus. Consequently, this service not only aids in effective message delivery but also plays a crucial role in cultivating a harmonious and interconnected community.

As Korean letters has gradually replaced Chinese characters, an increasing number of people in Korea have become less proficient in using Chinese letters flexibly. Consequently, it may be more effective to utilize proverbs that are both metaphorical and intuitive, rather than rely on idioms that infer their meanings through letters and pronunciations. Therefore, as our service utilizing idioms has proven its effectiveness, a project aimed at recommending situational proverbs is also expected to demonstrate significant impact not only in every conversations but also across various fields that require effective communication.

8. System Implementation Image (For Reference)