EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Philosophische Fakultät

Name: Ha

Vorname: Hong Seon

Matrikel-Nummer: 3978142

Adresse: Marbacher Str. 20
Steinheim a.d. Murr 71711

Hiermit versichere ich, die Arbeit mit dem Titel:

Classification of Korean Nouns as an Extension of Idioms

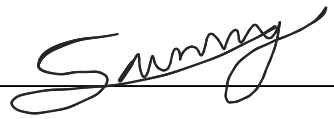
im Rahmen der Lehrveranstaltung ISCL

im Sommer-/Wintersemester WS 17/18 bei Dr. Çağrı Çöltekin

selbständig und nur mit den in der Arbeit angegebenen Hilfsmitteln verfasst zu haben.

Mir ist bekannt, dass ich alle schriftlichen Arbeiten, die ich im Verlauf meines Studiums als Studien- oder Prüfungsleistung einreiche, selbständig verfassen muss. Zitate sowie der Gebrauch von fremden Quellen und Hilfsmitteln müssen nach den Regeln wissenschaftlicher Dokumentation von mir eindeutig gekennzeichnet werden. Ich darf fremde Texte oder Textpassagen (auch aus dem Internet) nicht als meine eigenen ausgeben.

Ein Verstoß gegen diese Grundregeln wissenschaftlichen Arbeitens gilt als Täuschungs- bzw. Betrugsversuch und zieht entsprechende Konsequenzen nach sich. In jedem Fall wird die Leistung mit **„nicht ausreichend“ (5,0)** bewertet. In schwerwiegenden Fällen kann der Prüfungsausschuss den Kandidaten/die Kandidatin von der Erbringung weiterer Prüfungsleistungen ausschließen; vgl. hierzu die Prüfungsordnungen für die Bachelor-, Master-, Lehramts- bzw. Magisterstudiengänge.

Datum: 12.01.2018 Unterschrift: 

UNIVERSITY OF TÜBINGEN

BACHELOR THESIS

Classification of Korean Nouns as an Extension of Idioms

Author:
Hong Seon HA

Supervisor:
Dr. Çağrı ÇÖLTEKIN

A thesis submitted in fulfillment of the requirements for the degree of B.A.

in

ISCL

January 12, 2018

UNIVERSITY OF TüBINGEN

Abstract

ISCL

B.A.

Classification of Korean Nouns as an Extension of Idioms

by Hong Seon HA

Motivated by lack of previous research and similarities of English idioms and Korean nouns, the thesis attempts to adapt the English idiom classification task for the Korean noun classification purposes. The thesis works with the hypothesis that certain type of Korean nouns closely resemble English idioms, while the rest resemble non-idiomatic English phrases. Thus, the thesis is focused on the adaptation of English idiom classification task to evaluate the results, consequently proving or disproving the hypothesis. The specific classification tasks are done using softmax and recurrent neural network with word embeddings as methods and measures the accuracies of such methods. Due to the distribution of the data set related to real-life word distributions, the thesis deals with the majority class bias where nouns with Sino-Korean roots dominated the set in ratio of approximately 13:1 against the nouns of pure Korean roots. Through efforts discussed in the thesis, the classification tasks nullify the impact of such majority class bias in the end, further producing very good results in support of the hypothesis. While the results are satisfactory, the thesis goes over potential improvements that could be made and further applications of the experiments conducted in the thesis.

1. Introduction

This study is concerned with classifying specific Korean nouns as an extension of idioms. It attempts to investigate the hypothesis that certain Korean nouns are compositionally similar to idiomatic expressions, and therefore, idiom classification method can be adapted to work with Korean nouns. Thus, a successful execution of the classification task should, for example, separate the Korean nouns into two categories: a category for those resembling idiomatic English expressions and a category for those resembling non-idiomatic English expressions.

Idioms

Idiomatic expressions, or idioms, are used daily in regular written language and spoken conversations to improve the quality of such communications by shortening the words necessary to express certain ideas with its metaphoric meaning. Due to such nature of idioms, they present a compelling classification case - given a data set of English phrases, how could one identify the idioms from non-idiomatic expressions. An example of such is with the phrase *add fuel to the fire*, which has a commonly-used, figurative meaning of *to worsen an already difficult situation*. While the phrase can also have a literal meaning of adding fuel to a real fire, humans can intuitively observe the phrase in a context and classify the phrase as an idiom.

Such ability to distinguish idioms from the non-idiomatic expressions can have multiple merits. For human language speakers and learners, the mastery of idioms can greatly improve their conversational abilities, which can translate to many social merits (studies, businesses, relationships, and etc.). For machines, bigger benefits can be expected. For example, the ability to correctly classify and understand phrases with idioms could result in machine learning and artificial intelligence usages in i.e. translation systems, automatic responses, and word sense disambiguation, to name a few. These can then have wide applications in products such as translators, phones, home systems (i.e. Amazon Alexa), where such idiom processing is needed.

Given these merits, ample studies on the classification of idioms in English languages have been already conducted. For reasons such as the amount of data and academic/business interests; however, it is not the case in languages such as Korean. Nonetheless, Korean as a language still presents many interesting questions to investigate. In this vein, the thesis looks to perform idiom classification in Korean.

Korean as a language has gone through many historical changes. Ever since the 15th century when its writing system was first invented, it has gone through numerous changes over time, much like all other languages, dropping certain features and adapting others. In Korean, however, the cases of adaptation are especially significant. Having close economic and cultural ties to China, the neighboring country for

centuries, Korean language had adopted many features and words from the Chinese root. This was further instituted by the nobles in Korea, whom in elitism, continued to utilize Chinese writing system even after the creation of the Korean equivalent. Not only that, during the Japanese Imperialism in the early to mid 20th century, the Japanese officials put a country-wide suppression of the Korean language, leading to adaptations of many Japanese-rooted words as Japanese language became culturally embedded to the general population of the country. These exemplary cases as well as many others over the centuries, therefore, left Korean with interesting features, many of which are of Sino origin and permanently enrooted to the language itself.

Consequently, Korean is a very interesting yet unexplored language for machine learning in many different aspects. For the purpose of this thesis, however, only the Korean nouns are considered for their two distinctive features, which let Korean nouns be machine classifyiable data with two labels. The first of these two features is that Korean nouns are mostly comprised of nouns stemming from Sino-Korean (in layman's term, stemmed from Japanese and Chinese words and meaning) and "pure" Korean origins due to the historical context mentioned above. Korean nouns in general consist almost entirely of these two types of nouns with few exceptions of other foreign, borrowed words such as *컴퓨터* which is a direct translation of *computer* from English to Korean. Notably, due to this fact, many of Korean homonyms exist between two different nouns - one with Sino-Korean and one with pure Korean origin - i.e. '해' has meanings of 'damage' and 'sun' as homonyms. This feature allows the splitting of large set of Korean data into majorly significant two labels.

Secondly, Korean nouns are comprised of compound nouns and regular nouns all written without any spaces. An example of such composition is '시계,' which is translated to 'clock' in English. Broken down, the noun '시계' can be roughly understood as a compound of '시,' meaning 'time,' and '계,' meaning 'machine.' Additionally, to further expand on this idea, a 'watch' would be translated to '손목시계,' which is a compound of '손,' meaning 'hand', and '목', meaning 'neck,' combined together to mean 'wrist' with '시계', so literally a 'hand neck time machine' translated to 'watch'. As can be seen, opposed to the common English usage of spaced, less semantically understandable nouns, Korean nouns are created as compound nouns which often act similar to English phrases. These nouns, in extension, can be understood as a form of idiomatic expressions. As an idiomatic expression would express an unintuitive, deeper meaning, a Korean compound noun similarly conveys deeper word-relation context than just the noun. On the other hand, there are words such as '가위', which means 'scissor', and '버섯', which means 'mushroom' with no contextually-derived meaning. These are considered regular nouns, which are similar to regular, non-idiomatic English phrases in that they can be understood at face value. Furthermore, a closer observation of all the Korean nouns showcase that these regular nouns stem from pure-Korean root, while the compound nouns are of Sino-Korean roots.

Thus, when both of the features are examined in combination, Korean nouns illustrate an interesting phenomenon where nouns of the same lengths could be both regular and compound nouns (as with '시계' vs. '가위' and '버섯') that cannot be distinguished without any intuition. This would hypothetically then result in a classification problem very similar to that of an idiom classification in English. Just as one classifies the idiomatic expressions out from non-idiomatic expressions, one should also be able to classify the compound nouns (Sino-Korean nouns) from regular nouns (pure-Korean nouns). Therefore, borrowing this idea, the thesis will

examine the Korean nouns more thoroughly and adapt the English idiom classification for the Korean noun classification task.

For this purpose, Muzny and Zettlemoyer's paper was elected as the experimental basis. In the paper, Muzny et al. utilize English idiom Wiktionary dump to perform English idiom classification. Similarly, the thesis performs Korean noun classification using Korean noun Wiktionary dump. For the method of classification, the thesis opts for a simple softmax method and a more complex recurrent neural network deep learning method to analyze the performance differences and to draw more detailed conclusions about the Korean noun classificaiton task.

2. Process and Method

Formal Problem

Based on the introduction, the problem that the thesis would delve into is Korean noun classification using adjustments from English idiom classification method. This experimental problem can be formally written as such:

$$\forall n \in N, \text{ predict } \{y_n \mid y_n = p \ \vee \ s\},$$

where N = all nouns and y = binary label p = pure- or s = Sino-Korean. One noticeable adjustment from Munzy and Zettlemoyer is that there is no differentiation of noun phrase and definition. The two-variable pair is simply combined as a n. For example, the task at hand, given n = 친구, predicts whether or not n derives from Sino- or Pure-Korean root, or y. In order to accomplish the task, the machine is first trained on a train set of Korean nouns and their labels, gaining deeper knowledge on Korean characters and noun compositions as more data is fed. Once trained, the machine predicts through the test set, which includes the noun 친구 in the example, using the knowledge it has gained on Korean characters to guess the origin of 친구. These basic procedures and the results derived from the process are then evaluated and then repeated with addition of different features and variables.

Data Preprocessing

Following Muzny et al., the data processing was done on the latest Korean wiktionary dump collected on 20th August, 2017. Wiktionary dump, with proper processing, is a very helpful data dump especially for understudied languages such as Korean where computational resources are far scarce than i.e. English and Mandarin. The dump collected comes in xml format with many junk data. Thus, the dump is first pre-processed to obtain all Korean nouns and rewrite them to documents with individual data (such as origin, forms, and etc.). This was done using lxml and BeautifulSoup4 with python.

After the pre-processing, each document and its data is then utilized as filters for removing unqualifed nouns and false-positives from the pre-processing. All nouns that are used only in North Korean dialects, derive from other foreign origins (as exemplified before with 'computer'), are no longer used in Korean (i.e. nouns that include '.' used as a vowel in the ancient Korean language), and/or are Korean and Foreign character combinations (i.e. mixture of English and Korean, Chinese and Korean, Koreanized foreign words and Korean) are deleted from the documents. Ultimately, these filtered documents are then written in form of a list pair. The first

column of each row consists of the noun, and the second column of each row consists of the classification category (Sino-Korean or pure Korean root respectively marked as 'y' and 'n') of the noun.

In general, the data processing is carried out to ensure that no unique nouns stand out to affect the machine learning. Most of the processing was done based on the above filters while few had also been processed intuitively as for one reason or another i.e. some non-Korean nouns (i.e. Chinese characters) are falsely stored in the Korean wiktionary dump as Korean nouns and have to be manually evaluated and reduced.

In the end, the entire process allows the tuning of the 300mb+ data down to mere 400kb in the end, leaving approximately 28,000 noun-label pairs to the written list. This list is then separated into different train and test sets based on the experimented features.

Experiments

Given the formal problem, the thesis attempts to explore the classification problem in two major approaches. First approach is the simple softmax classification, in which only the basic configurations are made. There is no layers of embedding or any complex machine training algorithms, but a softmax function to learn and make the prediction. The second approach, in order to achieve more complex observation, is the Recurrent Neural Network. Through RNN, multiple different word embeddings layers and data sets are experimented on top of softmax classification further with more complicated configurations. These two separate approaches in comparison will help in identifying the complexity of the classification as well as learning about characteristics of the Korean nouns.

Softmax Classification

Softmax classification is a linear classification method that utilizes normalized probabilities as an end result. It is frequently used as a simple multiclass classification tool, often providing a good performance baseline for classification tasks. The result of the softmax classification can be sufficient for certain simple tasks; however, in most cases, more complex algorithms can achieve more advanced outcome. Below is the description of how the softmax classification is utilized for the experiment based on the online notes by Roelant.

Data and Model The idea of any classification task is to give train input to the system for it to learn by applying a chosen classification function on the input and using learned values to predict on the test input. The differences among classification tasks stem from the type of the given data and the method of categorization. Softmax classification, following this concept, is one of the simplest way in which the classification can be done.

The classification task can be broken down into two major steps: data processing and modeling. In data processing, the noun-label train and test sets are loaded in to be processed as machine-readable data. They are recoded so all word characters are represented by their own unique numbers (i.e. '친구' could be recoded to '3,1') and the labels into either 0 or 1. Then these nouns are split into smaller batches, ready to be put into a model. The recoded word characters are then truncated to prefix and suffix of certain lengths that are combined and used as features in the model.

In modeling, the model for softmax classification is set up such that for the training phase, train set batches would be taken in, calculated for its weights and biases, and transformed into logit values. These logit values would then result in losses after going through softmax cross-entropy function and optimized for minimization of the losses. In machine learning, these steps are considered to adjust or train the machine to achieve better prediction. These adjustments would continually be made throughout numerous epochs, optimally until the machine is evaluated to at its upper limitation in terms of accuracy. During the validation phase, these adjusted knowledge of the machine is tested based on the training by the test set batches, providing the accuracy that becomes the core measurement tool of the machine learning's success.

Configuration Softmax classification is an important part of the experiments in the thesis. Accordingly, there are multiple model configurations for the softmax classification that should be discussed for replication purposes. In the thesis, the softmax classification is largely used in three different ways. Firstly, it is used as a general classification tool with a simple setup. In this configuration, the batch size is set to 1,000 with each of the prefix and suffix lengths of the nouns to 3 based on experimentation. Furthermore, the learning rate is set to 0.0125 with its decay set to 0.90. The optimizer used is Adam Optimizer and the loss function is set to softmax cross-entropy with logits. Lastly, the number of epoch is set to 2,000 after multiple experimentations conveying that 1,000 is often not enough. Since the setup is to be simple and general, there is no complex regularization and normalization method utilized. In the second configuration, which is the configuration for the small experimentation set, the first configuration is mostly replicated with only the difference in batch size which is alternated to 256 to meet the difference in data set size. Lastly, in the third configuration, the normal data set configurations are taken and added an oversampling weight with input dropout of 0.65 (to combat model overfitting). This is done so by adding label weights to the each of the labels to minimize the major class bias before softmax cross-entropy function is applied. Other than such oversampling weight, all of other configuration settings of the two configurations are kept. Lastly, for all configurations, an early stop is set up to obtain peaking prediction accuracy most efficiently.

Recurrent Neural Network

Reccurent Neural Network (RNN) is an aritificial neural network deep learning approach that is often used in classification and recognition tasks. Compared to a softmax classification, RNN functions by adding a neural network algorithm layer before the softmax function is achieved. Through the process, it is able to process and obtain learnable information from arbitrary input sequences, thus, in an experiment that works with nouns with characters in certain orders, it is able to utilize the noun inputs for better results than softmax classification.

Word2Vec Word2vec is a tool used in machine-learning in which words and its relations are represented as word vectors. It is primarily used to provide word-relation context to data inputs that machines otherwise have hard time understanding. [4] For tasks such as speech recognition, all information required to successfully perform the task is encoded in the data as humans are able to perform such task themselves, given the same raw data. However, a machine learning system would treat words as distinct symbols rather than informative data, and therefore inputs such as

'시계' are represented as i.e. id0 and '컴퓨터' as id1. These representations by themselves are not enough for a sytem to extract data as humans intuitively do. The way they are represented, a system would only observe arbitrary symbols which provide no useful information regarding the relationships that may exist between the individual symbols. In turn, the model only utilizes so little of what it has learned about different inputs when it processes data of another input that to humans are related in word-relation context (think processing 'dogs' with data already given on 'cats' and 'hamsters'). Furthermore, since the thesis works on approximately 28,000 nouns, which is relatively low number of data set, using vector representations such as word2vec can help in giving more context to the training model to overcome the disadvantages that result from training on low amount of data set. The word2vec embedding is therefore used as an embedded layer in RNN approach in an attempt to test the hypothesis that there is word-relation context to Korean noun classification.

In order to further understand the impact of different word2vec embeddings in the neural network, multiple word2vec embeddings have been created with different configurations to test its implications. These embeddings are created on zero-restriction, filtering nouns of different lengths, training only on train set, and limiting the amount of total data set to be trained on.

Data and Model RNN utilizes the same data processing and modelling scheme as softmax classification. The difference between the two models exist as RNN does not utilize prefix-suffix features but a full length sequences of the word characters, and RNN also adds few more layers to the softmax model. Firstly, RNN utilizes the recoded word2vec embedding on top of recoded train and test data set which are all loaded into the model. Following, the recoded word2vec embedding is used within the model to create word embedding lookup. Lastly, the model is configured in the training phase so that the word embedding lookup and the data are taken in by the RNN cells to calculate hidden layer values for the use of logit calculations. These calculated logits follow the same procedure as the softmax model further on.

Configuration As a crucial part of the experiments in the thesis, RNN also consists of multiple model configurations to be discussed for replication purposes. RNN utilizes one general configuration that is used with minor tweaks throughout the experiments. The general configuration follows some identical configuration settings to softmax model with the learning rate set to 0.0125 with its decay set to 0.90, Adagrad Optimizer as the optimizer, and the loss function set to softmax cross-entropy with logits. On top of that, the model takes advantage of bidirectional GRU with cell states of size 200, input dropout rate of 0.85, and hidden dropout rate of 0.80 on for the training phase. The dropouts are necessary to not overfit the model. Additionally, the configurations are set up with 100 epochs and max timesteps of 20. For RNN, there are no prefix and suffix length configurations as the cells use dynamic sequence lengths. Based on this general configuration, the batch size of 256 is utilized for all experiments. Lastly, as with softmax model, some configurations are experimented with an addition of an oversampling weight. For all configurations, an early stop is set up as well to obtain peaking prediction accuracy most efficiently.

3. Results and Discussion

Data Preprocessing

The table below conveys detailed accounts of the preprocessed data. The data consists of noun phrases that are lengths of one to eleven. Most of these nouns, however, have the length of two to four characters per words/phrases. Out of the 28,804 total nouns, 2,096 of them originate from pure Korean roots while 26,708 of them originate from Sino-Korean or combination of Pure and Sino-Korean word phrases.

TABLE 3.1: Word Length Count of All Words on the List

n	Foreign	Pure	Sum
1	97	17	114
2	10674	726	11400
3	9403	760	10163
4	6053	416	6469
5	321	122	443
6	102	38	140
7	40	9	49
8	10	6	16
9	5	2	7
10	2	0	2
11	1	0	1
Sum	26708	2096	28804

The preprocessing produced quite a surprising result. As suggested by table 3.1, while restricted solely to the wiktionary dump data, Korean nouns mostly originate from Sino-Korean roots. While historical backgrounds suggest that this type of influence is evident, the scope of it is nonetheless intriguing. Still, it is unclear if the same trend can be portrayed from a bigger corpus such as that of a dictionary. Unfortunately, there is no real viable machine-readable source available for Korean corpus to further experiment this trend on.

Focusing on the data at hand, the ratio of the nouns rounded to approximately 1:13. Based on this ratio, it can easily be assumed that the experiments on the pure data set could have a majority class bias as a foreseeable problem. This occurs because since the probability that the class of the noun is Sino-Korean when the machine trains is significantly higher than the probability of the class being pure Korean, the machine could learn that it is more probable in almost all cases to simply choose the majority class than any other probabilities defined through the training.

So assuambly, any probability case that is lower than 1:13 would be disregarded in the actual prediction. In order to test and resolve such phenomenon, the main experiments will be conducted on all nouns in the list and on a restriction of 7,000 noun phrases for the Sino-Korean noun phrases for the train-test set to observe the phenomenon on different scales. Additionally, there would also be an oversampling approach on the entire corpora to balance out the bias via weights to observe the difference in results between this approach and the complete, unweighted list. Lastly, an experiment only on a small, restricted corpora of equal amount of nouns of both categories would be performed to ensure the natural effectiveness of the different classification methods.

In order to obtain quantitative measurements from the data sets and the classification methods, the above experimentations are done on both softmax classification and RNN classification of different configurations, which are further discussed in detail on upcoming segments. Overall, the validity of the thesis thus can be measured from the evaluation of the results originating from the different set ups encouraged by the data processing.

Baseline

Having a measurement guideline before getting into the experiment results is a good way to ensure that the analyses of the results are worthwhile and critical to useful discussions. One method in which this can be done is through the use of baseline, or basis for measurement. The baseline in computational experiments is usually the simplest and most intuitive method in which a task can be done. In the case of Korean noun classification, this would be picking Sino-Korean as the label every single time, which also happens to be the potential problem that has been previously discussed. Adhering to such baseline on all 28,804 nouns, for example, results in 92.72% prediction accuracy. Given this accuracy, any experiment results that are within a minor marginal difference in prediction accuracy can be said to have reached the baseline, most likely due to major class bias interfering with proper classification. It is also a good idea to keep in mind that even in this case, no prediction accuracy of exactly 92.72% would be achieved by the experiments since the data set used are divided into train and test sets in those cases, which the baseline didn't have to deal with. Furthermore, any prediction accuracy that is distinct to the baseline would suggest deviation from the major class bias. More on these results are to be discussed in later parts of the thesis.

TABLE 3.2: Baseline Accuracies of Different Data Sets

Data Set	# of Nouns	Baseline Accuracy
All Nouns	28,804	92.72
Limited Set	8370	83.63
Small, equalized Set	2600	50.00

The table 3.2 illustrates all the baselines necessary to know for evaluations of different results. As can be understood from the table, as the data set reaches closer to 1:1 ratio of the nouns of the two different labels, the baseline accuracy of the set gets closer to 50%.

Softmax Classification

With the baseline in mind, softmax classification is conducted on different variations of the data set and configurations, using the recoded data and softmax classification model. The table 3.3 conveys the results obtained from all softmax classification tasks.

TABLE 3.3: Softmax Classification

Data	Accuracy	Baseline
1. All Words	93.43	92.72
2. 7000 Sino-Korean	79.20	83.63
3. 2. & Without 1-Character Words	76.46	83.63
4. 3. & Without 2-Character Words	84.86	83.63
5. 1. & Oversampling	77.80	92.72

The table 3.3 shows prediction accuracies in all different ranges. With 1. data set, a very high result of 93.43% accuracy is achieved. Meanwhile, cutting the data down to size of approximately 9,000 with the reduction of the Sino-Korean noun phrases to the first 7,000 random phrases, or 2. data set, the resulting accuracy goes down to 79.20% accuracy. By additionally eliminating all one-worded nouns, or 3. data set, the accuracy adjusts to 76.46%. By further removing all two-worded noun phrases, or 4. data set, the accuracy of 84.86% is achieved. Lastly, by performing oversampling on the 1. data set, the accuracy of 77.80% is produced.

Primarily, these results are very insightful in understanding the data sets better. Firstly, the 93.43% accuracy that the 1. data set resulted in is very close to the previously mentioned baseline accuracy of the data set, while the oversampling on the same data set of 77.80% is approximately 15% lower than the baseline accuracy. This suggests that without any adjustments during the process, majority class bias is overfitting the model by 15%. The resulting accuracies of 2. and 3. data sets compared to their baseline show a few % decreases in accuracies. This conveys that by decreasing the ratio of the majority class to minority class, the system is able to eliminate the effect of majority class bias by a small but evident margin. The increased accuracy from the previous two data set results with 4. data set back to baseline accuracy, however, suggests that the majority class bias is again a factor with this data set. This is supported by the analysis of the preprocessed data that by such filter in data set, approximately 35% of the pure Korean nouns are removed solely from removing two-character nouns phrases. As the limit on the Sino-Korean words is still 7000 nouns, the filter only affects the number of pure Korean nouns, thus again amplifying the effect of majority class bias.

The above findings convey two major points. For one, it illustrates that oversampling is an effective method in reducing majority class bias while keeping the integrity of the data at hand. On the other hand, it also demonstrates that reduction and filters of nouns are not an effective method. This is because while the filters in some cases are effective in reducing majority class bias, wrong data set configurations can easily reintroduce the problem. On a larger scale, reducing the data set could introduce less to no problems, but especially with the small data set, it can very easily be affected by impacts previously not considered. Furthermore, while oversampling left the data set untouched, filters and removals method continued to cut down on the data set that was already deemed low in count. Therefore, oversampling seems to be the more reliable as well as accurate way of tuning the data

for precise and effective machine learning. Thus, the reliable outcome of softmax classification tasks suggests that, in general, this method is effectively able to predict at the accuracy of 77.80%.

Recurrent Neural Network

The results of RNN, compared to those of softmax classification, convey whether the word-relation context improves the general accuracy of the classification task. The table 3.4 illustrates all the results of such RNN experiments that have been acquired through the experimentation on mixture of two parameters: data restriction and different word2vecs.

TABLE 3.4: RNN Classification

Data Sets	TF	All	2+ Char Words	3+ Char Words	Split	Baseline
1. All Words	96.09	93.99	94.19	93.68		92.72
2. Limited	90.35	86.27	86.10	84.49	85.68	83.63
3. 2. & w.o 1-Char Words	88.90	84.54	86.55	81.64	86.72	83.63
4. 3. & w.o 2-Char Words	92.06	89.71	89.58	88.15		83.63
5. 1. & Oversampling		90.01	90.65			92.72

The table 3.4 describes the results of each experiment made on data sets (the row) using each of the word2vec embeddings (columns), wherever applicable. For the embeddings, 'TF' refers to the tensorflow-trained w2v, 'All' to training on all the nouns, and '2+ Char Words' to filtering out 1-character nouns from all nouns data set and training, '3+ Char Words' to filtering out both 1-character and 2-character nouns and training, and lastly 'Split' to leaving a space between every character of a noun on all nouns before training the w2v.

Very similar to the softmax classification results, the RNN results generally also convey that majority class bias is taking place. The baselines, which are 92.72% for all words data set, 76.96% for the data set limited to 7,000 Sino-Korean data set, and 83.63% for the limited set with further filtering of word lengths less than 2, are either reached or surpassed by the systems in almost of all situations, barring the oversampled experiments. Nonetheless, since multiple of these results also surpass the baselines by many percent, these results seem to indicate that more than just majority class bias is in play. Also, the oversampled RNN results show significant performance improvement of about 12% in accuracy compared to that of softmax classification.

RNN classification results leave a lot of room for discussion. Firstly, in comparison to the baselines, RNN showed a noticeable increase in accuracy across many experiments. Generally, this can be interpreted in two ways. On one hand, this signals that word-relation context of the word embeddings are helpful in overcoming the effects of majority class bias as results are higher in accuracy than those of the baselines in most experiments. On another, it could also be the product of additional data, not necessarily the word-relation context. This is predicted because the tensorflow-trained word2vec embeddings outperform pretrained, presumably more advanced, word2vec embeddings. Assuming that tensorflow's built-in word2vec has less functionality than a module created solely for the purpose of training words to vectors, such could be the case of tensorflow word2vec embeddings picking up additional noise from the data that deceptively contribute in the higher accuracy

during the classification. Given such evaluation, it is not hard to assume that there is a probable cause to be apprehensive of the significance of the observed data even for the word2vec-trained embeddings. Nonetheless, as resorting to complete skepticism would only deter any progress, the thesis leaves the point behind as a potential point for later inspection.

Comparing RNN results to the softmax results also seems to confirm that indeed word-relation context is positively affecting the performance of the Korean noun classification. While this is not a 100% concrete evidence of it, the comparison between the results of oversampled RNN experiments and that of softmax classification provides 12% increase in accuracy that can be only attributed to the influence of word embeddings as there exists no other distinctive difference in the two models except for the word embeddings. Hypothetically, one could still argue that this improvement is only an effect of word embeddings introducing noise to the data set. While this is not necessarily disproved in the thesis, on an intuitive level, it can be assumed that an increase of 12% in any weighted model cannot all be attributed to noises in the data. This is especially the case since all other configurations intended to observe the changes in accuracies based on them have proven insuccessful (i.e. splitting the words by spaces and creating different data sets and embeddings) in greatly adjusting the accuracy of the system. Thus, in terms of proving the hypothesis of the thesis, the Korean nouns do seem to have word-relation context, and a classification task of such requires relatively complex modelling method that utilizes the meanings of the words for robust performance.

Experimentation on Small, Even Data Set

TABLE 3.5: Experimentation

Data	Softmax	RNN - 2+ Char Words	Baseline
Split	N/A	74.02	50.00
Not Split	52.00*	74.71	50.00

The table 3.5 illsutrate the results of both softmax classification and RNN classification on a relatively smaller set of 2,600 nouns. The set has 1,300 Sino-Korean and pure Korean noun phrases each, which are separated into a train set and test set. The results convey an average of 52.00% for softmax classification and an average accuracy of 74.00% for RNN classification.

Such experiments on the development set was orchestrated to further test the relative influence of the word-relation context on the noun classification task as well as to observe for any potential anomaly in results. In the end, the remarkably strong performance of RNN on a relatively very small data set with 50% baseline accuracy compared to that of softmax classification, which barely overcame the baseline, further confirmed the previous assumptions that 1. softmax classification is a weak modelling method for such task, and 2. the word-relation context does exist in Korean nouns to help in classification of the nouns to their origins. It is, nonetheless, quite surprising that the boost in performance of RNN on classification over the softmax method is massive. Such phenomena is worth investigating with further research.

4. Conclusion

Based on the accumulated results and discussions, the hypothesis that certain Korean nouns behave similarly to English idioms and thus are machine classifiable seems very plausible. The evident improvement in accuracy that RNN classification has over softmax classification in all situations both with and without major class bias coincide with the premise of the hypothesis, given how the data was processed to be used for such purpose. However, as noted before, the pretrained word2vec embeddings' comparatively lower performance over the tensorflow-trained embedding suggests that, while difference between the two embeddings is not clearly identifiable, this difference could possibly be from unrelated noises. This assumption can be supported by the fact that pretrained word2vec would not have missed word relations if it existed, especially since the whole point of word2vec training was to make sure to emphasize on the word relations. If anything, the results of tensorflow's word embedding should have been lower than word2vec embeddings, if no additional noise existed. Nonetheless, while the thesis generated numerous experiments, of which many proved to be quite fruitless, the general outcome of the thesis is satisfactory. The thesis provides strong enough evidences that the Korean nouns are not random but built on certain structures based on meaning and word relations. It further presents useful model-building and fine-tuning recommendations as well as references to extend the research on Korean nouns.

With that said, the thesis is not perfect nor complete in its current status. While the experiments produced analyzable and discussable results, in hindsight, there are multiple improvements that could be made to further extend and add onto the experimentations of the thesis. One such improvement would be better measurement for the machine validations. The current thesis utilizes accuracy, which is quite generic, as its sole validation measurement tool. While accuracy as a measurement is helpful and insightful, often times in the thesis, there are moments when relying solely accuracy seems to lack. Such is the case especially since the experiments worked mostly with data sets that had majority class bias. Normally, when any system scores an accuracy of 90%, for example, it is deemed very robust. However, it is the case with majority class bias problems that accuracy of 90% only represents the skewed data distribution. If instead a measurement tools such as recall and precision or, similarly, confusion matrix were additionally used, all the measurements would have combined to bring more insights to what really is happening with the data. Unfortunately, instead, the thesis took a course of intuitively reducing such problems by other means of using the observed results. This, nonetheless, did prove to be successful but not truly satisfactory. For future experimentations, implementation of other measurements is highly recommended.

In addition, while the thesis kept on adapting and experimenting on new features and different data sets, it still seems to lack in its experimentation and evaluation of word2vec embeddings. Precisely, there have not been real effort to increase the size of word2vec input data size or to evaluate the effectiveness of the word2vec embeddings. While the results still indicate that word embeddings did prove to be functional in the classification problem, it is not clear whether the current results come anywhere near the most optimal outcome derived from optimal word embeddings or how precisely they contribute to RNN classification due to its size. It would therefore be a good idea in the future to potentially acquire i.e. a dictionary to learn word2vec embeddings from or, in the worst case, at least approach with synthetic data created from the pre-existing data sets to test out its optimal effectiveness.

In the end, however unfortunate, it must be pointed out that the amount of computational data for Korean nouns still lacks. This has unfortunately to do the most with not much research being done with Korean language as there is not as much merit compared to other popular languages. When and if such issue becomes less a problem as more data emerges, it would be interesting to see what kind of developments Korean NLP can take. However, it must be also noted that, as shown by the application of English idiom classification task for Korean noun classification task, there are quite possibly many other similar tasks that can be adapted from English NLP tasks that, given enough data, Korean NLP could naturally just follow. This is a basis for a hopeful sentiment that Korean NLP can improve without innovations but adaptations in the future.

In conclusion, the thesis adapted an English idiom classification task to Korean nouns in an attempt to learn and observe the validity of such classification task with Korean. While the thesis used relatively simple approaches and did not investigate deeply into the task, it still acquired acceptable results. With the outcome, the author hopes that more people in academia would be interested in conducting such adapted research in not only Korean but also other languages that lack much data set or any experiments.

Bibliography

[1] Grace Muzny and Luke Zettlemoyer *Automatic Idiom Identification in Wiktionary* 2013.

[2] Peter Roelants http://peterroelants.github.io/posts/neural_network_implementation_intermezzo02/ *Use of Softmax Function as a Classifier*

[3] https://colah.github.io/posts/2015-08-Understanding-LSTMs *Recurrent Neural Networks*

[4] https://www.tensorflow.org/tutorials/word2vec *Word Embeddings in Neural Networks*