Data Governance on Business/Data Dictionary using Machine Learning and Statistics

Philip Wootaek Shin
DataStreams Corporation
wtshin@datastreams.co.kr

Jinhee Lee
DataStreams Corporation
jinhlee@datastreams.co.kr

Seung Ho Hwang
DataStreams Corporation
shhwang@datastreams.co.kr

***Abstract*—This paper focuses on business/data dictionary to improve data governance for industries. Companies are expending considerable resources to govern their data. We have proposed applying different natural language method, viz., the statistical method and machine learning method. English terminologies are abbreviated because when they are saved in the database since the database has limited bytes for each column. English terminologies are abbreviated differently by different companies and even within a single company, the same word can be abbreviated differently. Therefore, companies are spending resources to maintain and standardize their data. Further, there are scenarios where two companies combine their data/business dictionary, in which case this integrated dictionary has to be re-standardized. We found a new method of applying these natural language processing approaches to the business/data dictionary for abbreviated English words which will help companies reduce the expenditure on standardization and improve the efficiency of data governance.**

***Index Terms*—Machine Learning, Natural Language Processing, Data Dictionary, Business Dictionary, Data Governance**

I. Introduction

As data becomes more and more important due to the evolution of machine learning, industries are willing to put in effort and resources to protect and govern the data they acquire. Data governance is the specification of decision rights and an accountability framework to ensure the appropriate behavior of data and analytics in valuation, creation consumption and control [1]. Specifically Business/Data Glossary or Business/Data Dictionary is the tool where businesses maintain and manage the business vocabulary that they use in field.

However, there have been issues where duplicate vocabularies are created and maintained on glossary/dictionaries which entails additional expenditure to business owners on consulting for removal of these duplicates and maintaining their data glossary [9] [10] [11]. In addition, in the case of mergers and acquisitions or when each subsidiary company has its own business glossary, and the parent company wants to merge all the business glossaries simultaneously, in order to merge different dictionaries removing duplicate words and unifying abbreviated style is crucial so that company can easily maintain a single data dictionary.

In this paper, a new method of applying these natural language processing approaches is proposed for business/data dictionary for abbreviated English words. This paper is organized as follows: Section 2 explains data governance, standard data dictionary and natural language processing. Section 3 illustrates integration of multiple data standards. Section 4 proposes a new method of applying natural language processing to data dictionary. Section 5 explains experimental results and Section 6 concludes this paper.

II. Background

The role of data science role has grown by over 650% since 2012 and is expected to create 121.5 million jobs by 2026 [1]. There are needs for businesses, corporations and enterprises to govern their business vocabularies, and there are researches done to help keep track of these business vocabularies. Further, multiple companies are providing these tools for these needs.

A. Data Governance

Data/Business Glossary has been studied by a few researchers, but generally this method was developed and distributed by companies. Informatica has a solution that uses business glossary for data governance and data quality [9]. IBM has InfoSphere Business Glossary which is a classification system for enterprise vocabulary [10]. Collibra's data governance tool also provides business glossary and data dictionary [11].

Metadata is often referred to as "data about data", which enhances the business and technical value of information stored in enterprise repositories such as operational data stores, information warehouses, and data marts [12]. For data governance it is crucial to keep track of this meta data. There are studies done on business vocabulary one of them using a questionnaire-based exploratory survey, which shows that business process management/information system development community is supportive of standardardization and integration of business vocabularies. However, existing standards like Semantics for Business Vocabulary and Rules(SBVR)are not widely known [13]. There is another study on Business Information modeling methodology and implemented Accurity Glossary [14]. There were limited information for Data glossary/Data dictionary since most of the concept and product

were used in companies that own solutions rather than the research fields.

B. Standardized Data Dictionary

Unlike any other country, the Korean Government's Ministry of Public Administration and Home Affairs has issued guidelines for database standardization and standardization of Data Dictionary for public institutions/Government Agencies. Standardized Terminology(Standardized term) is a business term that database users can communicate with. This term is created by standardized vocabulary that is defined by the institution/agency. Standardized Term Dictionary is the set of standardized terms and common standardized terms [15]. Since this standardized term dictionary concept is standard in Korean government, Korean industries, and corporations follow this standards when they employ this standardized dictionary.

C. Natural Language Processing(NLP)

Natural Language Processing(NLP) is the field where people try to train computers to understand human language. There have been different efforts to train computers in Thesaurus method, Statistical method, and Inference method [2]. Statistical approach uses corpus to do natural language processing. This method extracts important information from the corpus by the following procedure. First, the vocabulary is split from the corpus. After that, vocabulary is maintained with ID list and these vocabularies are vectorized to present it as a distributional representation. In the field of NLP, distributional hypothesis is defined as follows: vocabulary's meaning comes from the surrounding vocabularies which means that the context of the vocabulary comes from the surrounding vocabulary [3]. In order to keep track of the surrounding vocabulary from the corpus, co-occurrence matrix is created and the window size is defined to limit the number of vocabularies surrounding target vocabulary. For example if the window size is 1, it means the context is to be limited to the word that is just adjacent to the target word. The window size can vary but usually for business vocabulary, the vocabulary is constructed with less than 5 words.

For evaluating vector similarity, there are many methods such as Euclidean Distance Similarity and Cosine Similarity. Following previous work [4] [5], we use the commonly employed cosine similarity that is defined as follows:

$$CosineSimilarity(\mathrm{x}, \mathrm{y}) = \frac{\mathrm{x} \cdot \mathrm{y}}{||\mathrm{x}||\,||\mathrm{y}||} \tag{1}$$

There were efforts to improve this statistical approach using Pointwise Mutual Information(PMI). When using co-occurrence matrix the frequency of the word does not matter with in the size of the window. However, in sentences there were vocabulary like a, and the, which are not as important as other vocabularies to find the context of the word. PMI metric can help find these words and eliminate these components when finding context [6].

After the emergence of machine learning, there were many efforts to develop an inference-based approach for natural language processing. There were models like Word2Vec [4] [7], and GloVe [8] which tried to train computers in human language using machine learning layers and techniques.There are two models in Word2Vec which are the Continuous Bag of Words(CBOW) and Skip-gram model. CBOW model predicts the context of the target word using the surrounding words, and Skip-gram model uses the target word to predict the surrounding window of context words. The GloVe model uses statistics model and machine learning model.

III. Integration of Multiple Data Standards

When a conglomerate multinational company wants to expand its business, individual subsidiary companies operate on their own, manage their information and data separately, and create their local standards. However, it is crucial to keep track of global standards to exercise control over child companies. In case of mergers and acquisitions, the parent company, might want to migrate all the local standards into a global standard to take control over the child companies. There were generally two ways to make global standards between companies, viz., Centralized management and Distributed management [16]

Centralized management controls and manages standards of glossary and metadata by the centralization system, and in order to modify data, locals have to seek permission from the central authority. The distributed management system manages and controls glossary and metadata locally. When a conglomerate multinational company already has the centralized system, it is best for child companies to obey the rules and structure made by centralized system. This is a top-to-bottom approach which needs less effort since the child company just has to obey the rule. However, in the distributed system, child companies have their own rules and in order to have a unified dictionary or get rid of duplicate business vocabulary, there must be a process to gather all the information and get rid of duplicates. This is the bottom-to-top approach; since there are metadata and glossaries that were standardized locally, there needs to be process of combining and refactoring the data to conform to the global standards.

MetaStream [16] is a software tool created by our company that provides a data governance environment for defining and sharing standard terminology and the standard code to be applied by collecting table structure and code values for business systems. It also serves to provide reference data for data modeling and data standardization. MetaStream follows the guideline of Korea's Ministry of Public Administration and Home Affairs for data standardization and standardized term dictionary. However, since there were needs for integration of multiple data standards, we propose a technique to apply natural language processing to these different standards to integrate multiple standards, into one which is a bottom-to-top approach.

For standardizing multiple data dictionary we will use the statistical approach and machine learning approach. We propose a top 5 recommendation system that can list similar vocabulary from different dictionaries using two approaches. Before applying methods for data, data needs pre-processing,

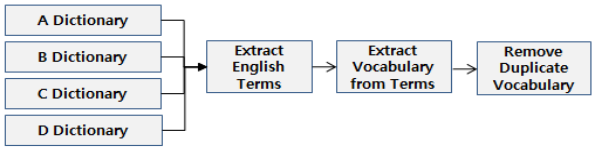

Fig. 1. Preprocessing of Data(Preprocessing unit)

as in Fig 1. First, from the different dictionaries of different companies, the English terminologies are extracted. English terminology consists of multiple vocabularies/words, and these dictionaries tend to be abbreviated since there are limits to the number of bytes stored for each column in the table/database. For example the word "Code" can be abbreviated as CD, CDE, CE, or even as CODE. For Korean government agencies and companies, since English terminology is the supplementary material, it is divided by "_"; for example, the business term "Market English Name" is abbreviated and saved in dictionary as "MRKT_ENG_NM". But generally, these abbreviated terms are used globally, since there are limited amount of storage in the column. Extracting this vocabulary from the business term is crucial because different businesses tend to abbreviate vocabulary as per their own rules, and there will be multiple abbreviated versions for the same word. In addition there are possibilities of finding the same abbreviated term from the different databases, making the removal of the duplicate vocabulary is an essential step for the future.

IV. Applying Natural Language Processing to Data Dictionary

Since Data/Business Dictionary also consists of English language(although abbreviated), it is suitable to apply natural language processing techniques to this dictionary to find the vocabulary similarity. This paper applied the statistical approach and machine learning approach as in Fig 2. and proposes a recommendation system.

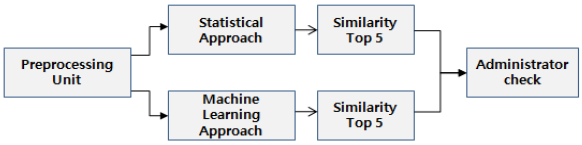

Fig. 2. Diagram for the recommendation system

There are three important methods of natural language processing, viz., thesaurus method, statistical method, and inference method [2]. For thesaurus method, there needs to be an encyclopedia that humans made for synonyms, on the basis of which considerable work and research have been done; however, for these rare abbreviated English vocabularies, there is not much to be done by the thesaurus method. Therefore this approach is excluded. After pre-processing through statistical and machine learning, the top 5 similar words are obtained for each method. We have chose top 5 recommendations, inspired by the top 5 accuracy(top 5 error rate) that is used as a metric for ImageNet [17]. For top 5 accuracy, the model's 5 highest probabilities matches the expected answer. After similarity Top 5 procedure, the administrator can collectively see what looked like similar words with two different methods and standardize the dictionary.

For dataset, we have used privately-owned datasets from 3 different companies(X,Y,Z). In order to exclude heterogeneity of dataset, we have limited the companies to the field of finance out of different fields like bank, credit card company, insurance company. Since these 3 companies are in the same field, the data dictionary of each company has similar business vocabulary/terms, standardized by their own rules. The dictionary of company X contains 28987 standardized terms and if this were to be divided into each word it has 3461 unique words, while company Y contains 14835 standardized terms with 2218 unique words. Company Z has 4684 business vocabulary/terms divided into 2303 unique words.

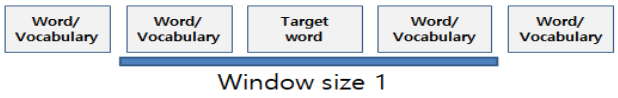

Fig. 3. Example of window size

For implementation both the approaches, window size was limited to size 1 which means for context it just uses one adjacent word next to the target word as in Fig 3. However, there might be a different variation that the user or administrator wants for words that are long; so, the window size can be modified as needed. In addition, we have implemented top 5 for the recommendation system but this too can be modified for different needs by switching the parameter size; for example, if you switch the parameter to 7, the recommendation system will show top 7 similar vocabularies.

A. Statistical Approach

The statistical approach follows the following procedure. First, find the target word vector from the list of vocabularies that were retrieved in the pre-processing unit. Next calculate all the cosine similarities of every word vector, and target word vector. Lastly print the top 5 based on cosine similarity calculation result. For implementation, there are two versions which were just the statistical approach that uses cosine similarity, and the method which contains pointwise mutual information [6], which reduces the redundancy of unrelated words and singular value decomposition, [18] which reduces the dimension of vectors to decrease the calculation when calculating cosine similarity for all the words.

We figured that pointwise mutual information and singular value decomposition did not play important roles in the abbreviated word, since for the business dictionary, depending on the product, the administrator has to approve the business terminology in order to be posted in the business dictionary; so point wise mutual information would not make a significant

difference. In addition, since the vocabularies that we collected were small compared to the natural language(human language), singular value decomposition didn't play a significant role in reducing time when calculating.

B. Machine Learning Approach

For the machine learning approach, there were different models for natural language processing but Word2Vec [4] [7] was chosen because of its good results in word embedding, word relatedness, and word analogy experiment [19]. For conventional machine learning training, there were training set, validation set, and test set. However, in our experiment, there were no validation, and data set, because had no prior permission been granted for the business terms/vocabulary, it would not appear in the data dictionary. So if it is not in the training set, it will not have a chance of appearing in the dictionary, instead it will spit out an error saying that it is unregistered vocabulary.

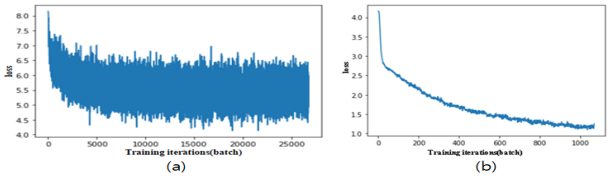

Fig. 4. The example of loss before fine tuning and after fine tuning, embedding layer, and negative sample applied. (a) shows that loss doesn't converge, but (b) shows that loss does converge with less training iteration

For training we have trained 3 different models, viz., Word2Vec with just company X data, Word2Vec with company X and Y data, and Word2Vec with company X, Y, and Z data. Without hyperparameter tuning, the loss does not converge. Loss means the score by the softmax layer for input and output layer and this probability is applied cross entropy error with the correct label. The embedding layer saves dense vector representation, which reduces the memory usage and unnecessary calculations [20]. In addition, negative sampling can reduce the softmax layer's calculation which would reduce the training time. Fig 4 shows the difference before fine tuning and after fine tuning with embedding and negative sampling [21] implemented.

V. Results

For evaluation, we mainly focused on the statistical approach with company X data, Word2Vec trained with company X data, statistical approach with company X and Y data, and Word2Vec trained with company X, Y data. For company Z data, the vocabulary and terms that they used have been too discrete and specific to their own company. Companies X and Y were relatively big ones, while company Z was really small and had its own unique rules and abbreviations.The result is shown in Fig 5.

```
[STAT1]YMD
 YM: 0.654021775683565
 DT: 0.5555176661189032
 NTRADE: 0.5232153002701556
 SERV: 0.5201294356033754
 CAROP: 0.49458104110733025
[STAT2]YMD
 YM: 0.6603583620489457
 DT: 0.5642055933082919
 SCTN: 0.5314058033612851
 NTRADE: 0.5232153002701556
 SERV: 0.5201294356033754
```

```
[INFER1] YMD
 RTUCH: 0.6982421875
 EXT: 0.68212890625
 DAYNUB: 0.66943359375
 WANT: 0.6630859375
 YM: 0.6435546875
[INFER2] YMD
 DAYNUB: 0.69580078125
 EXT: 0.67626953125
 WANT: 0.6728515625
 NUMT: 0.65771484375
 YY: 0.626953125
```

Fig. 5. The result of different methods: [STAT1] means it is the result of using 1 company's dataset with statistical approach, [STAT2] means it is the result of using 2 companies' dataset with statistical approach [INFER1] means it is the result of using 1 company's dataset with inference approach, [INFER2] means it is the result of using 2 companies' dataset with inference approach

A. Single Data Dictionary Analysis

The focus of this paper was to use these two methods in integrating multi data dictionary. However, this method could also be used in the single data dictionary case since a lot of Korean companies spend large amounts of money on consultation to standardize and remove the duplicate vocabulary/similar terms. Company X dataset has 28987 standardized business terms divided into 3461 unique words/vocabularies. It is best to analyze all the possible vocabularies covering the 3461 words and find the similarity of each one to standardize; however, that process does not make much difference if we analyze the whole thing, because that can be done by the consultant if the company pays for standardizing this dictionary. We have randomly selected 10% of the vocabularies. The result are shown in Fig 6.

```
SELL: sell
[STAT1]SELL
 BUY: 0.9642017308846954
 RPY: 0.8808367937491628
 LIM: 0.8743115569846881
 CNTC: 0.8650075771560007
 FACE: 0.8545202908676623
[INFER1] SELL
 BUY: 0.814453125
 SNS: 0.7880859375
 OPNT: 0.76708984375
 BWRRT: 0.7509765625
 DD3: 0.75
BUY: buy RPY:repay LIM:limit CNTC:contact FACE: face
BUY:buy SNS:sell OPNT:opening market BWRRT: bond with warrant DD3::3 Day Deal

New: new
[STAT1]NEW
 TDAY: 0.7023347505512263
 BUY: 0.6847815150771549
 RPY: 0.6793002214912655
 SELL: 0.6761580260902932
 P1D: 0.6726376984053399
[INFER1] NEW
 WEB: 0.78759765625
 NW: 0.75048828125
 ODRY: 0.72705078125
 BKAY: 0.71240234375
 CGRP: 0.70654296875
TDAY:The Day BUY:buy, RPY:repay SELL:sell P1D: past 1 Day
WEB:website NW:new ODRY:ordinarily BKAY:breakaway CGRP: C group

YM: year month
[STAT1]YM
 YMD: 0.654021775683565
 SERV: 0.5537540963113932
 CAROP: 0.4813595583346241
 NTRADE: 0.4700576571717318?
 STDEV: 0.4551395897444117
[INFER1] YM
 STRT: 0.841796875
 CAROP: 0.81103515625
 END: 0.80419921875
 YY: 0.79736328125
 DST: 0.75048828125
YMD:year month day SERV:serve CAROP:car operation NTRADE:nontrading STDEV:standard deviation
STRT:start CAROP:car operation END:end YY: year, DST:distance
```

Fig. 6. Example of vocabulary that matched similar/same vocabulary with in the same dictionary. For convenience, we have written what each abbreviated word corresponds to in the original English vocabulary after each result.

With in the same dictionary there were identical words with different abbreviations like SELL/SNS and NEW/NW with machine learning method and there were similar vocabulary like YM, which is similar to YMD using the statistical method. Within a single data dictionary, application of natural language processing can be useful for standardization of the dictionary.

B. Multiple Data Dictionary Analysis

For analyzing multiple data dictionary, the following preprocessing is needed, as indicated in Fig 1 and Fig 2 which was mentioned in previous chapters. Company X has 28987 business terms and divided in to 3461 unique words, while Company Y has 14835 standardized business terms and divided into 2218 unique vocabularies. When these two were combined it yielded 5679 unique vocabularies which were abbreviated but out of these vocabularies we could find 518 that were abbreviated in the same way for the same vocabulary. For this result, we randomly selected 10% of the vocabularies and after the process, we selectively chose the words that we thought we could find in other Business dictionary. The results are shown in Fig 7

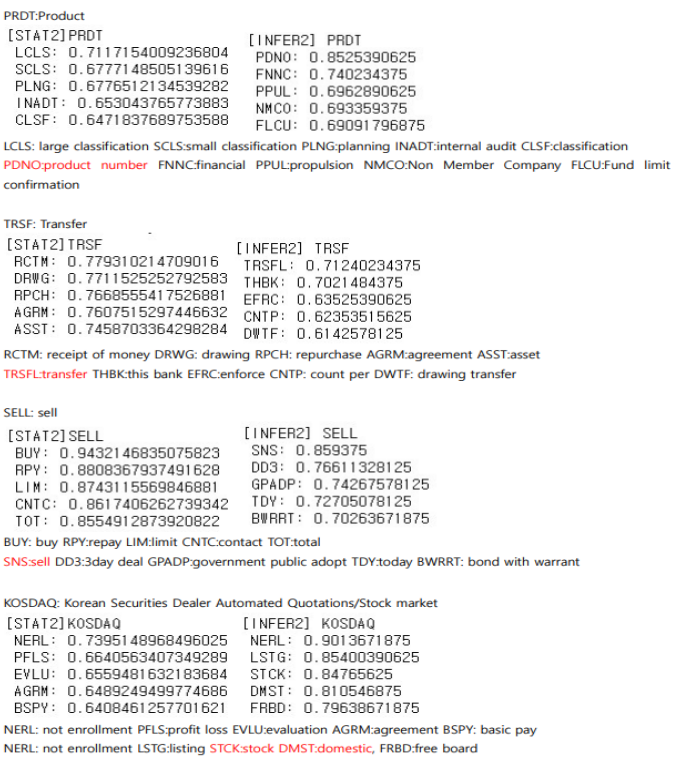

Fig. 7. Example of vocabularies that matched the similar/same vocabulary with in the integrated dictionary. For convenience, we have written what each abbreviated word corresponds to in the original English vocabulary after each result.

Fig 7 illustrates that there are different scenarios for the integrated data dictionary. We could still find SELL/SNS that could be found in the same dictionary which means the dictionary wasn't standardized well. In addition we could find the same exact word "transfer" abbreviated in two different ways TRSF/TRSFL in different data dictionaries. Some interesting results could be found: PRDT which product in company X dictionary and PRNO was product number in company Y dictionary. This means that there are high probabilities of different abbreviation for Product number such as PRDTNO instead of PRNO since that abbreviated vocabulary can be abbreviated differently or subdivided into a smaller abbreviated word. In addition, KOSDAQ which stands for Korean Securities Dealer Automated Quotation which is the stock market in Korea, we could find some interesting facts that similar words were found as STCK and DMST which are abbreviated word for stock and domestic. Since KOSDAQ is a domestic stock market for Korea, it was interesting and reasonable to find machine learning suitable for analyzing the integrated data dictionary.

VI. Conclusion

This paper proposed a new method of applying natural language processing methods to data/business dictionary. We expect this approach to be the stepping stone in applying machine learning method to data governance system and reduce the resources required for standardizing the data dictionary.

For natural language processing there are numerous test metrics like relatedness, analogy, categorization, selectional preference, etc [19], because in human language, there are grammar rules, and lots of researches were done since the invention of computers to make computers understand human language. However , there is not much research done on this field of abbreviated business words. In addition even though the dictionaries that were from the same field, there was heterogeneity among this dictionaries because they were not from the same company or subsidiaries. So depending on the area of standardization each company wants to focus on, different results can emerge, in which our studies can be helpful by using the statistical and machine learning approaches, database administrator can get recommendations on similar vocabulary to easily implement standardization efficiently and economically. This research can be solidified with more collections of different datasets, but the reality is that the actual datasets were privately owned by each company. For future research we will extend the focus to the business term itself rather than the single vocabulary, and domain classification system that will help data governance.

References

[1] Paul C. Hershey, "Data analytics implications" IEEE Potentials July/August 2018
[2] Sycho Goki(2019). Deep Learning from Scratch 2. Republic of Korea Hanbit Media
[3] Z.Harris, "Distributinal structure". Word, 10(23): pp. 146-162, 1954
[4] Thomas Mikolov, Ilya Sustkever, Kai Chen, Greg S Corado, and Jeff Dean, 2013 Distributed representations of words and phrases and their compositionality. In NIPS, pages 3111-3119
[5] Ronan Collobert, Jason Weston, Leon Bottou, Michael Karlen, Koray Kavukcujoglu and Pavel Kuksa. 2011. Natural language processing from scratch. JLMR, 12:2493-2537
[6] [11] K. W. Church, P. Hanks, "Word association norms, mutual information, and lexicography," Compututational linguistic Vol. 16, No. 1, pp 22- 29, 1990
[7] T. Mikolov, K. Chen, G. Corrado, J. Dean, "Efficient estimation of word representations in vector space,." arXiv prepreint arXiv:1301.3781, 2013.

[8] J. Pennington, R. Socher, C. Manning, “GloVe: Global Vectors for Word Representation,” Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Quatar, pp.1532-1543, Oct. 2014.
[9] Infomatica. ”Infomatica Business Glossary Overview” Internet:https://docs.informatica.com/data-quality-and-governance/data-quality/10-2/business-glossary-guide/user-and-role-administration.html[Nov. 14, 2019]
[10] IBM.”IBM InfoSphere Business Glossary” https://www.ibm.com/support/knowledgecenter/es/SSZJPZ-9.1.0/com.ibm.swg.im.iis.productization.iisinfsv.overview.doc/topics/c-bg-and-bga.html[Nov. 14, 2019]
[11] Collibra ”Collibra Data Governance ” https://www.collibra.com/data-governance[Nov. 14, 2019]
[12] Chaki S. (2015) Glossary of Terms. In: Enterprise Information Management in Practice. Apress, Berkeley, CA
[13] Kapocius,Skersys and Butleris. (2014) “The Need for Business Vocabularies in BPM or ISD Related Activities: Survey Based Study”
[14] Priebe, Markus “Business Information Modeling:A Methodology for Data-Intensive Projects, Data Scienece and Big Data Governanace”
[15] Ministry of Public Administration and Home Affairs, “Guidelines for Database Standardization in Public Institutions”, Ministry of Government Administration and Home Affairs No. 2015-26, 2015. (Korean)
[16] DataStreams ”Data Governance - Metadata Management ” URL: http://datastreams.co.kr/en/sub/prd/governance/metadata.asp [Nov 18, 2019]
[17] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li and L. Fei-Fei, ImageNet: A Large-Scale Hierarchical Image Database. IEEE Computer Vision and Pattern Recognition (CVPR), 2009.
[18] G. H. Golub, C. Reinsch, ”Singular value decomposition and least squares solutions,” Numer. Math., vol. 14, pp. 403-420, 1970.
[19] T. Schnabel, I. Labutov, D. Mimno, T. Joachims, “Evaluation methods for unsupervised word embeddings,” Proceeding of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, pp. 298-307, 2015
[20] O. Levy, Y. Goldberg, I. Dagan, “Improving Distributional Similarity with Lessons Learned from Word Embeddings,” Transaction of the Association for Computational Linguistics, Vol. 3 pp 211- 225, 2015
[21] O. Levy, Y. Goldberg, I. Dagan, “Improving Distributional Similarity with Lessons Learned from Word Embeddings,” Transaction of the Association for Computational Linguistics, Vol. 3 pp 211- 225, 2015