Toward Latent Knowledge Extraction Based on the Correlation of Heterogeneous Text Data Related to Space System Development

Kenji Mori
Japan Aerospace Exploration Agency
Tsukuba, Ibaraki, Japan
mori.kenji@jaxa.jp

Naoko Okubo
Japan Aerospace Exploration Agency
Tsukuba, Ibaraki, Japan
okubo.naoko@jaxa.jp

Yasushi Ueda
Japan Aerospace Exploration Agency
Tsukuba, Ibaraki, Japan
ueda.yasushi@jaxa.jp

Masafumi Katahira
Japan Aerospace Exploration Agency
Tsukuba, Ibaraki, Japan
katahira.masafumi@jaxa.jp

Toshiyuki Amagasa
University of Tsukuba
Tsukuba, Ibaraki, Japan
amagasa@cs.tsukuba.ac.jp

Abstract

This paper highlights the importance of careful selection of appropriate NLP tasks or techniques to derive value from past documents and improve the requirement engineering process. As a case study, an experience about introducing NLP techniques to find the lack of requirements by using heterogeneous documents are shown. Using word similarity is one of the ways to determine the relevance between two documents though, the result of proposed scheme in finding meaningfully related pairs of document and further investigation shows that word similarity is not able to solve our problem. In our experimental results, CNN (convolutional neural network) model could estimate the relevance the best compare to other trial models.

1 Introduction

Recently, deep learning has led to remarkable improvements in Natural language processing (NLP) research [You17]. As a result, not surprisingly, many industries encouraged to introduce NLP to improve internal engineering process using various types of document resources which contain potential value [Fal13]. However, careful selection of appropriate NLP tasks or techniques is quite important to derive values for one's problem.

In: M. Sabetzadeh, A. Vogelsang, S. Abualhaija, M. Borg, F. Dalpiaz, M. Daneva, N. Fernández, X. Franch, D. Fucci, V. Gervasi, E. Groen, R. Guizzardi, A. Herrmann, J. Horkoff, L. Mich, A. Perini, A. Susi (eds.): Joint Proceedings of REFSQ-2020 Workshops, Doctoral Symposium, Live Studies Track, and Poster Track, Pisa, Italy, 24-03-2020, published at http://ceur-ws.org

Table 1: Features of Design Documents

Content	Design bases including analysis results. Specification of functions. Must satisfy requirements for the target component.
Features	High abstraction especially in early development phase
Average number of words	215.9 Japanese words
Quantity	470,000 words

Table 2: Features of Anomaly Reports

Content	Factual information such as what happened, its cause, and countermeasure report. Interpretation of issues to make it reusable in other products.
Features	Collected across products, companies and writers. There can be a lack of information.
Average number of words	231.7 Japanese words
Quantity	41,923 items

In this paper, we show our experience about introducing NLP techniques to find the lack of requirements by using heterogeneous documents maintained in JAXA (Japan Aerospace Exploration Agency). It is vital to make the best use of the past documents due to the characteristics of space systems, i.e., the products are not mass-produced, and their development life cycle is very long, e.g., 20 years. Here, anomaly reports are selected to explore the insufficient requirement by searching the relevance between the development documents, thereby contributing to preventing previously experienced anomalies. In JAXA, requirements for a software component are organized and technically documented as a design document which is reviewed circumspectly. Hence, the design document is thought to be an appropriate target to confirm the adequateness of their requirements from the viewpoints of both the contents and the importance of the process phase gate. More precisely, if we can present past anomalies that are related to a design document with high precision, we think that users may recognize potential errors in the design. Besides, it may contribute to training novice designers. From these perspectives, we think finding valuable anomaly reports which provide unknown, latent, or forgotten knowledge about the target software component is same as estimating the goodness of correlation between anomaly reports and design documents.

We first explain the features of two target documents, namely, past anomaly reports, and design documents. Tables 1 and 2 show the features of each document. Let us look into the characteristics of anomaly reports. They are created by various workers and are stored in a web-based system and used to manage the status of anomalies and, more importantly, to prevent recurrence of previously experienced anomalies. Thereby in general, the description tends to be specific, but they often contain ambiguity caused by human, e.g., spelling variations and abbreviations. Also, the reports are collected across a wide range of products (launch vehicles, spacecraft, ground systems, etc.) and causes (deterioration, incorrect operation, logical errors, etc.). Consequently, we observe a significant discrepancy between the anomaly reports and the design documents in various aspects. On the other hand, the design documents are generated during development processes and referred to in technical reviews through phased project planning.[1] As describe before, the review of design documents seems to be an important phase gate to ensure requirement quality, thus finding related past anomaly reports is beneficial.

In our first challenge, it was difficult to find the relevance of two documents with simple keyword matching (Section 2). The underlying assumption with the keyword matching was that relevance of anomaly reports and development documents can be explained by the similarity, especially the similarity of keywords. This result highlights the importance of selecting appropriate NLP task or techniques which match with the feature of one's problem, also the feature of language resource.

The contribution of this paper is twofold. First, a simple scheme to associate heterogeneous documents (development documents and anomaly reports) using word embedding and convolutional neural network (CNN) is proposed and its results show the capability of the proposed scheme in finding meaningfully related pairs of document portions (Section 3 and 4). Second, the importance of selecting the appropriate NLP task is highlighted since word similarity is not able to solve our problem (Section 5).

[1]The actual samples of the design documents and anomaly reports were introduced at Appendix A.

Table 3: Training parameters for word2vec

Training model	skip-gram
Window size	10
Vector size	200
Epochs	5
Negative sampling	10
Learning rate	0.025
Batches	1,000

2 Preliminaries

The goal of this work is to find related anomaly reports for a design document, thereby contributing to preventing previously experienced anomalies. Two preliminaries would like to be introduced in this section, one is using keyword matching technique, and another based on document similarity comparison. Although they used some statistical processing for accumulated documents, were not kinds of supervised learning approach. Thereby they can be applied with reasonable cost, but their performance was insufficient.

2.1 Keyword Matching Approach

At first, we adopted a simple keywords matching approach to detect valuable correlations between anomaly reports and design documents. Keyword matching was executed using five keywords extracted from the target document. Okapi BM25[Rob09], which can quantify word importance from a document based on the frequency of word appearance, was used to select the keywords. Three examineers evaluated the results about twenty design documents. The top twenty anomaly reports extracted by the similarities based on keyword matching were evaluated respectively in sense of correlation to design document. As a result, the precision rate was 8.25%, which is insufficient. It revealed that the approach seems not to have the capability to achieve the goal.

There were two issues related to such low precision. One is the homograph problem, for example sometimes "wheel" means parts of ground vehicles, and in other situations it means a component of a reaction wheel, which is used to control the attitude of satellites. This issue leads to anomaly reports with different contexts that are likely to be found with high similarity. Another is the spelling variations problem. Although "Star Tracker," "star tracker," "STT," and "Star Tracker" in Japanese have completely the same meaning, keyword matching cannot treat them as the same. Although preparing a dictionary for name collation may be a solution, the cost will not be reasonable.

2.2 Sentence Similarity Approach Using Word Embedding

As the second preliminary, we tried to estimate the correlation using a similarity of two documents derived from a word embedding technique. The cosine similarity of the document feature vectors is used to score the correlation, and the document feature vector d was simply defined as below.

$$d = \frac{1}{n} \sum_{i=0}^{n} v_i \tag{1}$$

v_i is an embedding vector of a word that appeared in the document, and n is the number of words. A famous method word2vec[Tom13] was adopted to obtain appropriate word embedding vectors. All the anomaly reports and Wikipedia topics related to space system development were used for the training. There were 760,337 documents including 101,991 vocabularies, and other parameters were listed in Table.3.

The evaluation was conducted using the same design documents and examineers as in Subsection2.1. The precision rate is 19.3% which is better than the result of keyword matching approach. Although it was not also enough performance, the two problems described at previous subsection was mitigated.

At first, the top five similarities of the keyword matching approach includes 44.0% misguided reports by the influence of the homograph problem, on the other hand the error rate of this approach was 10.0%. We think the advance was from the pros of the document feature vector which can contain the whole document not only keywords. The basic potency of the word embedding which can quantify semantically similar words as similar vectors would also contribute to make the bad effect of the spelling variations problem small.

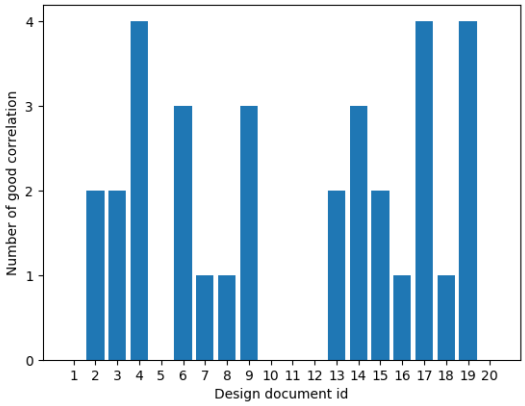

Figure 1: The result of the approach using word embedding

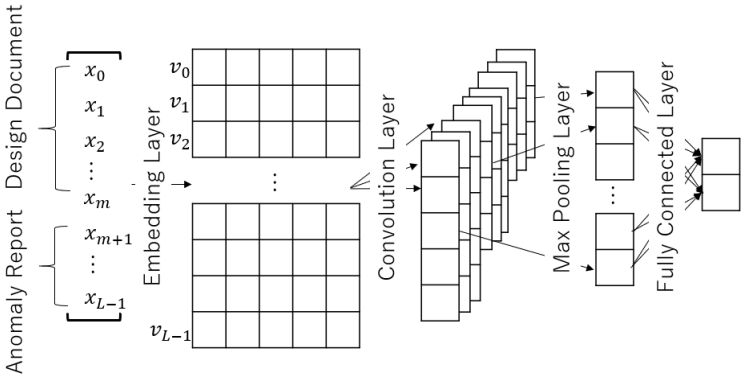

Figure 2: Proposed model

Additionally, there is a wide range of variations about the number of good correlations depending on the query design document. Figure 1 shows the number of good anomaly reports for each design document in the top five similarity. Looking at the details of the results, the design documents with high precision have two properties. First is the contents which was described about a single topic. For example the design document indexed 17 which has only the definition of the reference time in the module is an sample. On the other hand, index number 20 is about the porting of a function which has been implemented as hardware to software, thereby there are words concerning to both hardware and software. Second is about devices which have discriminative character. The device named star tracker which is an optical sensor to capture the designated planets to estimate the pose of the satellite is one of an example. No.19 has both properties.

3 Correlation Estimation Model

Based on the results and consideration of preliminary experiments, we used CNN to estimate the correlation between a design document and an anomaly report. The effectiveness of CNN for the document classification tasks has been reported[Kim14][Ye15]. Figure 2 shows our model.

3.1 Dataset

We prepared a dataset to train and evaluate our model. One record consists of a design document and an anomaly report, and it is annotated by three examiners who have similar expertise. They judged whether checking the anomaly report is useful in preventing design flaws against the design document. Items which were judged valuable by two or more examiners were treated as positive.

3.2 CNN Architecture

This subsection describes the details of the CNN architecture. We formulated correlation estimation as a classification task that classifies whether a document pair is valuable. It was designed to be simple and have few parameters. Complex, large architecture models are thought to be difficult to apply to our problem, because a large scale dataset would be required in order to train them, but our dataset was not so. As described in Subsection 3.1, it was prepared with limited resources from full scratch.

The input for CNN is the list of words included in a design document and an anomaly report. A document was divided into words by MeCab[Tak04], which is a popular morphological analysis tool for Japanese text.

The popular word embedding layer was applied to the input layer, since it was showed there is not an obvious relationship between the quality of word vectors and one of a downstream task[Gla16]. The length of the input word list is defined as L and x_i means the i-th word's one-hot-vector representation. The vocabulary file was prepared in advance to create one-hot-vectors. By using an embedding layer, the i-th one-hot-vector becomes a word embedding vector $v_i \in \mathbb{R}^{N_{word}}$. N_{word} is a predefined dimension of the word embedding vector. When the length of the input word list is less than the defined length which is $m + 1$ for the design document in Figure 2, zero fill vectors are placed in the vacant area.

The architecture has several size filters in the convolution layer. The filter size can be defined as $\mathbb{R}^{N_{word} \times L_w}$, L_w means the window size of a filter. The filters will learn to be able to extract word sequence patterns which are essential to being classified. The max pooling layer gets its strongest signal from the output of the convolution layers. Although the position information of the word sequence pattern is lost by this operation, we considered it acceptable to classify the patterns because it is an element which varies by writer, the category of articles, and so on. After that the fully connected layer learns the combination of carried signals and judges whether an input document pair is valuable or not. The softmax layer normalizes and outputs two signals, one means the pair is valuable, another means not.

Although this architecture does not directly compare the two document features to judge their correlation, the relation was thought to be acquired during the learning phase. In other words, each design document has both a related anomaly report and not related one in the training dataset, therefore the CNN will learn to extract features from them. In addition, from our viewpoint, the functional similarity of words is thought not to be equal to the value of correlation in this case. This is the reason why the architecture was adopted. Although there might be some cases in which valuable correlation items have a high functional similarity, it is not always true. We will discuss this in Section 5 with an experiment.

4 Evaluation

As an evaluation for our model, we conducted the comparative experiment between the model estimations and the judgments by the examiners.

The dataset described at Subsection 3.1 consists of 5,000 pairs. There are 10 anomaly reports for each of 500 design documents. The design documents were selected from technical documents for about four different earth orbiters. The technical domain is about attitude and orbit control system of a satellites. There are 882 positive and 4,118 negative samples.

4.1 Training Parameters

Parameters for the training of the CNN are shown in Table 4. In the training step, 882 negative samples were randomly picked from the population of negative samples to avoid incorrect learning due to the imbalance of positives and negatives. As the validation set, 15 positive and negative cases were also selected at random. So in the training phase, 1,734 samples with the same number of positives and negatives samples were used, and 30 samples are used to validate the generalization performance.

4.2 Results

Forty pairs of design documents and anomaly reports were prepared to evaluate the model. The pairs in the set consist of 4 design documents and 10 anomaly reports each. They were completely separate from training data for both design documents and anomaly reports. The examiner was also different from the person who annotated the training set. There were 17 positive samples and 23 negative ones.

Accuracy, precision, recall rate, F measure and mean average precision (MAP)@5 are measured to evaluate the model performance. These scores are the average of 10 training / evaluation trials. When measuring the

Table 4: Training parameters for CNN

Length of design documents	400
Length of anomaly reports	431
The number of vocabulary	18,274
Dimension of embedding layer	200
Window sizes of convolution filters	3, 4, 5
Number of convolution filters	128
Epochs	200
Batch size	30
Learning rate	0.0001
Optimization	Adam

Table 5: Evaluation score of the CNN

Accuracy	Precision rate	Recall rate	F measure	MAP@5
77.5%	79.2 %	64.7%	71.0%	80.0%

MAP@5 score, the results with only one design document of 4 are collected first. By sorting the 10 items based on the softmax layer output value to arrange results similar to the recommended ones.

Table 5 shows the results and their distributions as a boxplot in Figure 3. In the cases in where pairs with their softmax output value is more than 0.9 are treated as the model's recommendations, the precision rate becomes 89.0%.

The recall rate is slightly worse than other scores and the accuracy score for only positive samples is 64.7% and the one for negatives is 87.0%. From these results, we can say that the model is useful in removing such anomaly reports that are useless. However, if we think about using our model as a part of recommender system, it is desirable if it recommends anomaly records from different perspectives while avoiding recommending similar ones. We recognize that the *recommendation diversification* is an important issue, and address it as a part of our future work.

The proposed model worked more powerfully than the statistical approaches described in Section 2. The bad effects from homograph and spelling variation problems were also decreased. Homograph problems will be slightly improved by using whole documents in the estimate. The embedding layer will learn to treat spelling variations the same in training as same as the word embedding approach. In this case, the vector in the embedding layer for "STT" and the one for "Star Tracker" in Japanese is quite similar. It is 14th from the top when sorting all 18,274 word vectors based on the cosine similarity against the "STT" word vector.

5 Comparison with the Similarity Based Approach

The CNN-based architecture is used in our model to estimate the values of correlation among design documents and anomaly reports, which are annotated by examiners. On the other hand, the similarity among documents is often used as an important feature with keyword matching approaches or so. However our underlying assumption is that it will not work sufficiently for our task because the correlation among our target documents is not simple from the viewpoint of knowledge. To confirm this, an experiment using a similarity estimation approach was done as described in this section.

In the experiment, the similarity is estimated using a Siamese network[Bro93]. It is a well known model that can handle the similarity of two chunks of data, and recently the performance of the framework in learning the similarity of sentences has been reported[Mue16]. Hence the similarity can be said to be used as the basis to judge the value of the correlation, if the Siamese network can estimate them with high accuracy.

This network consists of two parts. One is the feature extraction part which is a neural network, another is the similarity measuring part which compares two feature vectors. The weights of the feature extraction network are trained to reduce the error between the feature distance and the supervised distance. For example in a case of similarity measuring, the positive sample distance as a supervised signal will become 1 and a negative one will become 0 or -1 when using cosine similarity as the distance measuring method.

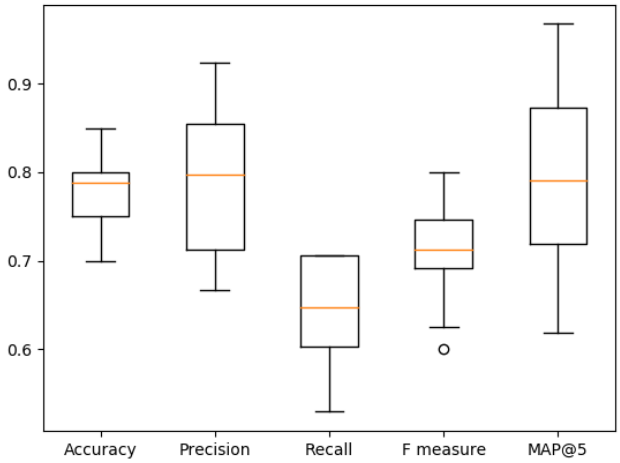

Figure 3: Evaluation results

Table 6: Evaluation score of the Siamese network

Accuracy	Precision rate	Recall rate	F measure	MAP@5
35.3%	33.0 %	54.7%	40.9%	21.3%

5.1 Siamese Network Model

Figure 4 showed the overview of our Siamese network model. The CNN described in Subsection 3.2 was adopted as the feature extraction part. There are two slight modifications. One is the output dimension of the fully connected layer - it was modified to 256. The second is the length of input words list. It was changed to 431. The distance was calculated based on the dot product, therefore the distance of a valuable pair is 1 as a supervised signal.

5.2 Evaluation of Similarity Based Model

This subsection describes the evaluation of the Siamese network model. It was trained using the same dataset described in Subsection 3.1. The evaluation set and other parameters are also the same as Subsection 4.1 and 4.2. However the number of epochs was doubled to 400, because the training was not going well.

The results were listed in Table 6. They are significantly lower than our proposed model(Table 5). Figure 5 shows the loss and accuracy value shift during a training run. The loss cannot be reduced enough to achieve sufficient accuracy even for the training set. It revealed that the valuable correlation judgements by examiners should have different attributes from the documents similarity. These results validate our underlying assumption.

6 Conclusion

We explored appropriate techniques to estimate the value of correlation between a design document and an anomaly report in terms of whether it is valuable knowledge to find the lack of requirements that raise the recurrence of past anomalies. As a result, the CNN based model which can learn non-linear relationship worked effectively. In our evaluation experiments, the performance of the model achieved a 71.0% F measure and 80.0% as MAP@5. This is better performance than other approaches described at Section 2 and Section 5. Especially the performance of Siamese network shows that the correlation, which would like to be estimated, seem to be a different feature from the similarity of two documents. This would be important information to modify the model aiming to improve the performance or to find new models.

As for future works, we are planning to use the model in an actual space system development process. The trial in the process will be able to extract improvements to use efficiently in the actual development. Although in the evaluation the satellite attitude control systems were focused, challenges to other technical areas is necessary

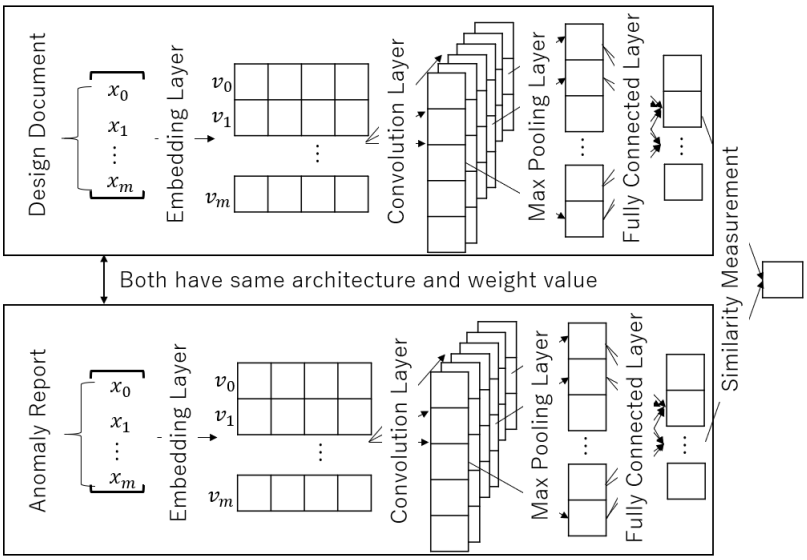

Figure 4: Siamese network model

to bring the benefit widely in our organization. We will not only apply additional annotated data but will also use transfer learning or more sophisticated neural language models.

Appendix A Sample of documents

Actual samples of the design document and the anomaly record are introduced here. The sample design document is about the input generated from the star tracker for attitude control calculations. Here is the translation result from Japanese to English.

> The star tracker captures and tracks several visible stars and outputs their position, star coordinates and star luminosity in the field of view. The obtained star coordinates are subjected to star identification processing and attitude determination processing, and the results are output as attitude quaternions. (Number of words: 45)

Two contrasting estimation results were shown in Table 7. The correlation column means the relevance against the design document. The first content shows the potential failure mode which may occurs when using the star tracker for the attitude control calculation. Although the content of the second example refers to the optical component, the main topic is about the calculation of field of view outside the context of attitude control calculations.

References

[You17] Young, Tom et al. Recent Trends in Deep Learning Based Natural Language Processing [Review Article]. *IEEE Computational Intelligence Magazine*, 13.3 (2018): 55–75. Crossref. Web.

[Fal13] Falessi, Davide et al. Automated classification of NASA anomalies using natural language processing techniques. *2013 IEEE International Symposium on Software Reliability Engineering Workshops, ISSREW 2013*, pp. 5-6, nov 2013.

[Rob09] Robertson, Stephen et al. The Probabilistic Relevance Framework: BM25 and Beyond. *Foundations and Trends® in Information Retrieval 3 (4)*, pp. 333-389, 2009.

[Tom13] Tomas, Mikolov et al. Distributed Representations of Words and Phrases and their Compositionality. *Advances in Neural Information Processing Systems.*, oct 2013.

[Kim14] Kim, Yoon et al. Convolutional Neural Networks for Sentence Classification. *Foundations and Trends in Information Retrieval*, pp. 1746-1751, oct 2014.

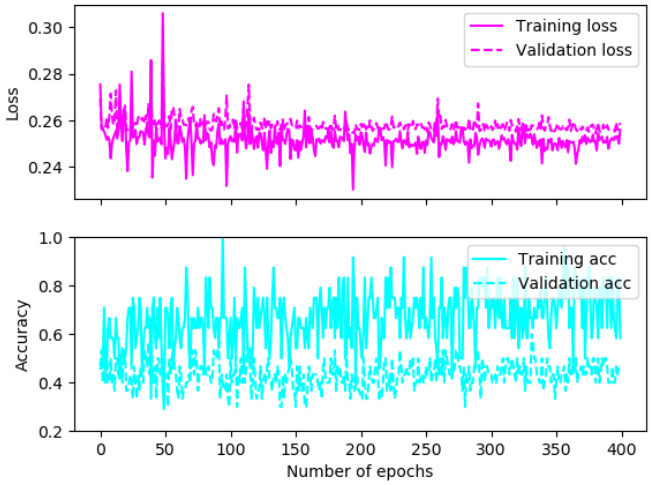

Figure 5: Loss and accuracy shift during a training

Table 7: Features of Anomaly Reports

Content	Correlation
The reaction wheel operated with a larger control amount than expected. The data from the STT was rejected, the attitude determination process was performed using the attitude data including the error, and STT was unable to identify stars, and repeated acquisition and tracking. During that time, the input and update of the attitude quaternion are not performed, the attitude is determined based on the data including the error obtained from the other sensors, and the attitude angle recognized by the attitude control software and the actual attitude angle are shifted. Data from STT was rejected even if STT returned since the gap between data from STT and calculated attitude angle get larger than the threshold. (Number of words: 115)	✓
When capturing a target object with an optical camera, the field of view range was calculated. However, the pose of the target object and the own pose were set in reverse. Therefore, in conjunction with the gradient of the object pose, field of view for the optical camera inclined as well. Then, the angle of field of view get different between + X-axis direction and the -X direction. (Number of words: 70)	×

[Ye15] Ye, Zhang et al. A Sensitivity Analysis of (and Practitioners' Guide to) Convolutional Neural Networks for Sentence Classification. *arXiv preprint arXiv:1510.03820*, 2015.

[Tak04] Taku, Kudo et al. Applying Conditional Random Fields to Japanese Morphological Analysis *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP-2004)*, pp. 89-96, may 2004.

[Gla16] Gladkova, Anna et al. Intrinsic Evaluations of Word Embeddings: What Can We Do Better?. *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pp. 36-42, aug 2016.

[Bro93] Bromley, Jane et al. Signature Verification Using a "Siamese" Time Delay Neural Network. *Proceedings of the 6th International Conference on Neural Information Processing Systems*, pp. 737-744, 8 1993.

[Mue16] Mueller, Jonas et al. Siamese Recurrent Architectures for Learning Sentence Similarity. *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, pp. 2786-2792, 2016.