# Effect of Preprocessing for Distributed Representations: Case Study of Japanese Radiology Reports

Taro Tada Kazuhide Yamamoto

Nagaoka University of Technology, Nagaoka, Japan

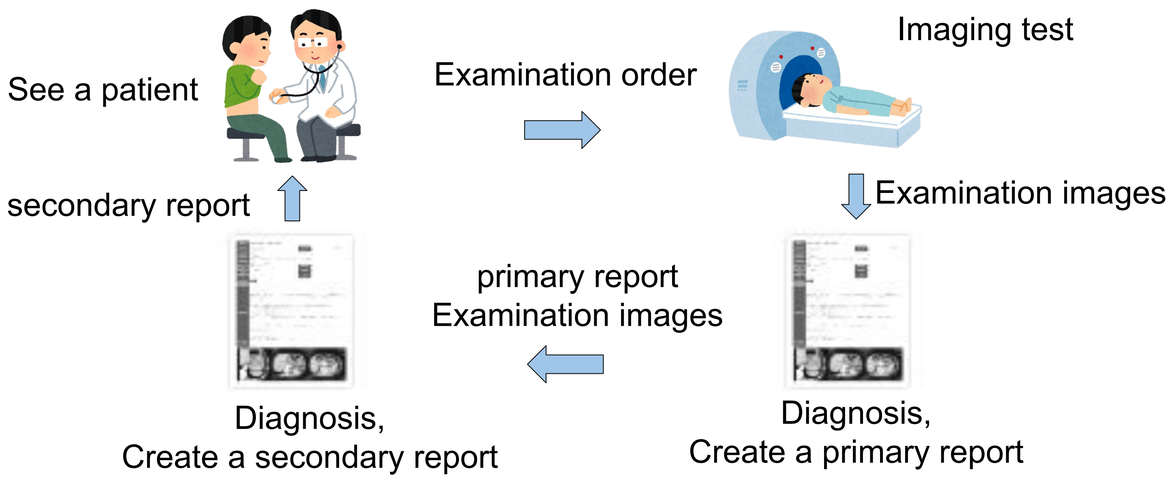
## ① Background

### Motivation

- **Radiology report creation is a heavy burden for doctors.**[1]
  →Demands similar documents retrieval system.
- **Not many studies using distributed representations in Japanese medical document retrieval.**
  →Unclear what kind of pre-processing is better for Japanese medical documents.

**Investigated on preprocessing of Japanese medical documents for distributed representation methods.**

### Japanese radiology reports




- **Examination order documents:**
  Order for examination by the physician who saw the patient.
- **Primary radiology reports:**
  Created by looking at examination images.
- **Secondary radiology reports:**
  Created by looking at examination images and a primary report.

## ② Data

Train data：Documents of 181,875 exams
Evaluation data： Secondary radiology reports of 80 exams
（8 disease categories x 10 documents）

Statistical information of Documents

| Type of document | Average document length (char.) |
|---|---|
| examination order | 99.1 |
| primary radiology report | 299.3 |
| secondary radiology report | 332.3 |

### Evaluation data

- **No medical domain dataset available for evaluation in Japanese.**
- **Built with reference to Okamoto et al.**[2]

**Many examination reports have multiple diseases.**
→ It is difficult to determine the degree of similarity between documents even if manually.
→ Simplify the problem as a first step for the tasks, select documents so that diseases are independent between categories.

**8 disease categories：**
Alzheimer's disease, lung cancer, myocardial infarction, fatty liver, disc herniation, medial collateral ligament injury, elbow fracture, Achilles tendon injury

**Calculate a cosine similarity between each document of evaluation data and the remaining 79 documents.**
- Confirm categories of top 9 similar documents for every document.
- Same category or not.

## ③ Experiments・Results

### Effect of word segmentation granularity

Example of segmentation results with MeCab[3]

| dictionary | segmentation results |
|---|---|
| example1：右肺上葉切除術後 | (After upper right lung lobectomy) |
| UniDic | 右肺 上葉 切除 術後 |
| IPADic,IPADic-NEologd | 右 肺 上 葉 切除 術 後 |
| MANBYO Dictionary, ComeJisyo | 右 肺上葉切除 術 後 |
| example2：陳旧性心筋梗塞 | (old myocardial infarction) |
| UniDic | 陳旧 性 心筋 梗塞 |
| IPADic,IPADic-NEologd | 陳 旧 性 心筋梗塞 |
| MANBYO Dictionary, ComeJisyo | 陳旧性心筋梗塞 |

Effect of word segmentation with MeCab

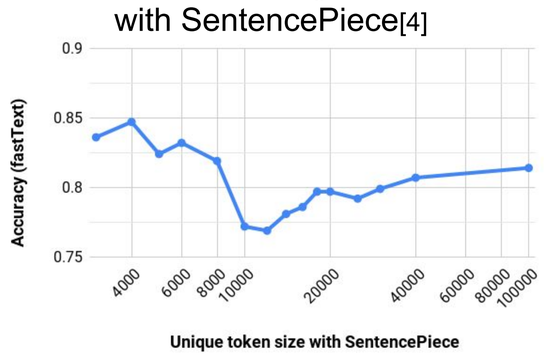
| segmentation method | fastText | Doc2Vec |
|---|---|---|
| character segmentation | 0.685 | 0.799 |
| UniDic | 0.708 | 0.894 |
| IPADic | 0.704 | 0.908 |
| ipadic-neologd | 0.663 | 0.896 |
| MANBYO Dictionary[1] | 0.644 | 0.828 |
| ComeJisyo[2] | 0.613 | 0.869 |

NOTE: fastText and Doc2Vec accuracies are not comparable.

Statistical information of text after segmentation

| Segmentation method | Average document length (word) | Average token length (character) |
|---|---|---|
| | secondary reports | secondary reports |
| character segmentation | 332.3 | 1 |
| UniDic | 237.0 | 1.40 |
| IPADic | 234.9 | 1.41 |
| IPADic-neologd | 226.7 | 1.47 |
| MANBYO Dictionary | 219.3 | 1.52 |
| ComeJisyo | 211.9 | 1.57 |
| SentencePiece(4000) | 147.6 | 2.25 |

Effect of word segmentation with SentencePiece[4]




### Effects of removing frequent phrases

In radiology reports
-Description of all parts displayed on the images.
(Negative findings that are not the main test part tend to describe body parts continuously)
-Negative descriptions also appear in the main test body.
-Greetings
→ **Experimented simple cleaning**

1. **Check frequently** used phrases from training data.
2. **Select sentences** that are likely to be noisy and negative phrases by manually.
3. **Remove selected phrases** from evaluation data.

Effect of removing phrases from evaluation data

| Method | fastText | Doc2Vec | Avg. length (words) |
|---|---|---|---|
| IPADic without removing | 0.704 | 0.908 | 151.4 |
| IPADic with removing | 0.781 | 0.922 | 130.7 |

1.MANBYO Dictionary(http://sociocom.jp/ ~data/2018-manbyo/index.html)
2.ComeJisyo(https://ja.osdn.net/projects/comedic/)

Reference:
[1]Kyoko Makino et al., Development and Evaluation of a Diagnostic Documentation Support System using Knowledge Processing,Transactions of the Japanese Society for Artificial Intelligence,Vol.23 No.6 pp.463-472 2008
[2]Kazuya Okamoto et al., Context-based Retrieval System for Similar Medical Practice Documents,Transactions of Japanese Society for Medical and Biological Engineering, Transactions of Japanese Society for Medical and Biological Engineering,Vol.49 No.6 pp.199-206 2006
[3]Taku Kudo et al., Applying Conditional Random Fields to Japanese Morphological Analysis, Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, pp.230-237 2004
[4]Taku Kudo et al., SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing, Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (System Demonstrations), pp.66-71 2018

Acknowledgment:
Thanks to Y's Reading inc. for giving incisive comments and sharing anonymized text data of examination documents.

## ④ Conclusion

- Word segmentation using MeCab with dictionaries are more accurate when it is finer. In addition, the word segmentation using SentencePiece is also more accurate when the number of unique tokens is as small as 4000.

→**The frequency of tokens appearing in the learning data is considered important.**

- In cleaning the evaluation data, it was possible to easily remove the factors that unnecessarily increase the similarity between the documents in the interpretation report by using a high-frequency phrase in the learning data.

→**Cleaning demands little domain knowledge and manual operation.**