DATA PAPER

RCMR 280k: Refined Corpus for Move Recognition Based on PubMed Abstracts

Jie Li[1,2], Gaihong Yu[1†], Zhixiong Zhang[1,2,3]

[1]National Science Library, Chinese Academy of Sciences, Beijing 100190, China

[2]Department of Library, Information and Archives Management, School of Economics and Management, University of Chinese Academy of Sciences, Beijing 100190, China

[3]Key Laboratory of New Publishing and Knowledge Services for Scholarly Journals, Beijing 100190, China

Keywords: Refined corpus; Move recognition; Sequential sentence classification; Corpus construction; Corpus analysis.

Citation: Li, J., Yu, G.H., Zhang, Z.X. RCMR 280k: Refined Corpus for Move Recognition Based on PubMed Abstracts. Data Intelligence 5(3), 511-536 (2023). doi: 10.1162/dint_a_00214
Received: November 15, 2022; Revised: March 10, 2023; Accepted: April 10, 2023

ABSTRACT

Existing datasets for move recognition, such as PubMed 200k RCT, exhibit several problems that significantly impact recognition performance, especially for Background and Objective labels. In order to improve the move recognition performance, we introduce a method and construct a refined corpus based on PubMed, named RCMR 280k. This corpus comprises approximately 280,000 structured abstracts, totaling 3,386,008 sentences, each sentence is labeled with one of five categories: Background, Objective, Method, Result, or Conclusion. We also construct a subset of RCMR, named RCMR_RCT, corresponding to medical subdomain of RCTs. We conduct comparison experiments using our RCMR, RCMR_RCT with PubMed 380k and PubMed 200k RCT, respectively. The best results, obtained using the MSMBERT model, show that: (1) our RCMR outperforms PubMed 380k by 0.82%, while our RCMR_RCT outperforms PubMed 200k RCT by 9.35%; (2) compared with PubMed 380k, our corpus achieve better improvement on the Results and Conclusions categories, with average F1 performance improves 1% and 0.82%, respectively; (3) compared with PubMed 200k RCT, our corpus significantly improves the performance in the Background and Objective categories, with average F1 scores improves 28.31% and 37.22%, respectively. To the best of our knowledge, our RCMR is among the rarely high-quality, resource-rich refined PubMed corpora

† Corresponding author: Gaihong Yu (E-mail: yugh@mail.las.ac.cn; ORCID:0000-0003-1301-2871).

available. Our work in this paper has been applied in the SciAIEngine, which is openly accessible for researchers to conduct move recognition task.

1. INTRODUCTION

Moves of research papers refer to the linguistic knowledge units that researchers used to express their writing purposes in academic exchanges, such as research purposes, methods, results, conclusions etc. Move recognition can help researchers quickly grasp the main points of research papers, and it is useful for various text-mining tasks such as information extraction, information retrieval and automatic summarization.

Move recognition methods has also developed from traditional machine learning methods, such as Native Bayes (NB)[1], Conditional Random Fields (CRF)[2], Support Vector Machine (SVM)[3, 4], Logic Regression (LR)[5], to neural networks methods, such as LSTM[6], Attention-BiLSTM[7], CNN[8], HSLN-RNN[9]. Especially the BERT[10] pre-trained language model, which has achieved better results in classification algorithms.

The corpus is a very important basis in move recognition task. The quality and quantity of the corpus can directly affect performance of move recognition model. The richer and more accurate the corpus, the better the performance of move recognition model; on the contrary, the poorer and more distorted the corpus, the worse the effect of the annotation model. Therefore, constructing a refined move corpus is an important task for move recognition an also increasing more and more attention by researchers.

In recent years, scholars have proposed different move recognition corpus of abstracts and full texts papers in different disciplines, and published a series of structured move corpora, such as NICTA-PIBOSO[11], PubMed 200k RCT[12], CSAbstruct dataset[13], PubMed 380k[14], Emerald 110k[15], Fund project abstract move data[16]. In our previous works[17], we use PubMed 200k RCT and its subset PubMed 20k RCT to train deep learning classifiers to achieve sentence classification tasks, we find that the accuracy of OBJECTIVE and BACKGROUND are always lower than the other three move sentences.

In order to improve sentence classifier performance, we find that the existing datasets for classifying sentences has obviously problems: such as wrong labels between the semantic of the sentences and the labels, short/meaningless fragments caused by incorrect sentence segmentation methods etc. Our main contributions are summarized as follows:

(1) Analysis the Problems of PubMed 200k RCT.
(2) Propose feasible structured abstract refining methods.
(3) Construct a Refined Corpus for Move Recognition in general medical filed based on PubMed Research Papers, called RCMR.
(4) Construct a subset of RCMR in RCT filed, called RCMR_RCT.
(5) Conduct experiments to validate the corpus quality of RCMR_RCT and RCMR.

The following content is arranged as followed. Section 1 introduces the meaning of move recognition and the lack of high-quality data sets in this field. Section 2 reviews the current data sets in move recognition and analysis the problem in the current data sets. Section 3 gives the construction process of our refined data set RCMR. Section 4 conducted dataset evaluation experiments to test its validity and efficiency. Finally, we conclude our work in Section 5.

2. RELATED WORK

2.1 Existing Structured Abstracts Datasets for Sentence Classification

In 2011, Kim et al [11] developed NICTA-PIBOSO, a corpus consisting of 1,000 manually annotated medical abstracts from MEDLINE. Specifically, it includes 500 abstracts from the areas of traumatic brain injury and spinal cord injury, as well as another 500 abstracts from various medical issues, such as "Systematic Review of the Literature Regarding the Diagnosis of Sleep Apnea." These abstracts were classified into six specified medical categories: Population, Intervention, Background, Outcome, Study Design, and Other.

In 2017, Frank presented PubMed-200k RCT[12] (subset is PubMed-20k).This dataset consists of approximately 200,000 abstracts of Randomized Controlled Trials(RCT), totaling 2.3 million sentences. Each sentence of each abstract is labeled with their role in the abstract using one of the following classes: background, objective, method, result, conclusion. PubMed 20k [12] is a subset of PubMed-200k, it is a smaller dataset, which is chosen from the 200k abstracts by taking the most recently published ones, its format is as the same as PubMed-200k RCT.

In 2018, Cohan and Beltagy[13, 18] constructed CSAbstruct dataset collected from the Semantic Scholar corpus[13, 18], Which has 2,189 computer science abstracts (15 thousands sentences) and each sentence is manually annotated according to their rhetorical roles in the abstract. Similar to the PUBMED-RCT categories. Each sentence label in CSAbstruct is with one of 5 categories: BACKGROUND, OBJECTIVE, METHOD, RESULT, OTHER.

In 2018, Moura et al[14] proposed PubMed 380k, which was downloaded from PubMed/Medline, including 387,705 abstracts and completely contain five categories: Background, Objectives, Methods, Results, Conclusions.

In 2019, Stead et al [15] published a multidisciplinary dataset for abstract sentence classification, called Emerald 110k, The dataset contains sentences retrieved from multi-disciplinary non-biomedical journal abstracts, including 103,457 structured abstract (1,050,397 sentences). Each sentence is classified as belonging to one of the following heading classes: Purpose, Design/Methodology/Approach, Findings, Originality/value, Social implication, Practical implications, Research limitations/implications.

In 2022, Zhao et al[16] construct a structured Chinese fund project abstract dataset based on rules and deep learning method, in order to provide data support for Chinese fund project abstracts move recognition

system. This corpus includes 4 move labels: Background and problems, objectives and tasks, method content, value and significance. Table 1 gives the Overview of existing datasets for sentence classification.

Table 1. Overview of existing datasets for sentence classification.

Dataset	# Size	Domain	Manual	Year
PubMed RCT 200k	200k	RCT	N	2017
PubMed RCT 20k	20k	RCT	N	2017
NICTA-PIBOSO	1k	Medical	Y	2011
CSAbstruct	2.2k	Computer Science	Y	2018
PubMed 380k	380k	Medical	N	2018
Emerald 110k	110k	Multi	N	2019
Fund-Project-data	9k	Multi	N	2022

2.2 Problem Analysis of Existing Sentence Classification Datasets

Previous studies on identifying sections in scientific abstracts have some limitations, include: 1) smaller quantity: The available datasets are not large enough to be applicable for training big sentence classification models effectively; 2) the scope of existing datasets is not extended to the entire domain of medical. For example, NICTA-PIBOSO extract and manually annotate only 1,000 abstracts from medical subdomains, although it has good quality, the quantity of this corpus is not enough to train the classification model, especially deep classification model, like Bert [13]. CSAbstruct has 2,189 abstracts from the field of Computer Science, compared with PubMed 200k RCT, CSAbstruct exhibits a wider variety of writing styles, since its abstracts are not written with an explicit structural template, and it has a limited quantity data. Fund-Project-data is fund project which is not applicable for medical abstract sentence classification. Emerald 110k mainly concerns non-biomedical abstracts, however, the biomedical disciplines abstracts have predominately not been included in dataset.

There are two corpora related to our constructed corpus: PubMed 380k[14] and PubMed 200k RCT. We describe them in Section 2.3 and Section 2.4. In Section 2.3, we discuss the PubMed 380k corpus, which consists of 380,000 abstracts from a wide range of biomedical publications. This corpus serves as a significant resource for researchers studying the biomedical domain, and it provides an extensive dataset for developing move recognition methods. In Section 2.4, we focus on the PubMed 200k RCT[12] corpus, which includes approximately 200,000 randomized controlled trial (RCT) abstracts. And this dataset is specifically tailored for research on RCT abstracts and offers a more concentrated resource for studying RCT-related topics. It is particularly useful for analyzing the move structure and content of RCT abstracts. By comparing and contrasting these two corpora, we can better understand the strengths and weaknesses of each and determine how our constructed corpus improves upon or complements the existing resources.

2.3 Problem Analysis of PubMed 380k

PubMed 380k[14] is a corpus of annotated abstracts extracted from PUBMED/MEDLINE, which opted to follow the two steps construction strategy of Hirohata et al. In the first extraction step, this corpus builders

collect 3,326,605 structured abstracts. All collected abstracts were written in English and annotated with predefined section labels provided by the abstract's authors. In many cases, these predefined section labels of segment fall into typical categories 'BACKGROUND, OBJECTIVE, METHODS, RESULTS, CONCLUSIONS', in other words, there are various labels in the corpus, since the set of labels used in each abstract may vary depending on its publication source, thus in order to avoid a large number of different labels, they collapse similar section labels based on the mapping provided by the U.S. National Library of Medicine (NLM). After that, all section labels were mapped into five rhetorical roles, namely: Background, Objective, Methods, Results, and Conclusions; In a second step, they select only the abstracts containing the five mentioned rhetorical roles in its abstracts' structure.

After this filtering, the corpus was reduced to 387,705 abstracts. The resulting corpus was segmented into sentences using the Punkt Sentence Tokenizer from the NLTK library, totaling 4,924,037 sentences. Each sentence was annotated with a label corresponding to the rhetorical role of the section to which the sentence belongs, i.e., all sentences in Methods sections were labeled as Methods, and so forth.

In PubMed 380k, each abstract has completely five rhetorical roles. But, of which there are 69,135 abstracts (account for 17.83%) not appear in a sequence, like B-O-M-R-C, in which there are 12,942 abstracts end with Background; 4,298 abstracts end with Methods; 14,094 abstracts begin with Methods. These abstracts are compliance with standardized IMRAD formats (INTRODUCTION, METHODS, RESULTS, and DISCUSSION) for structured abstracts that reflect the process of scientific discovery[19], which have been defined for original research studies, review articles and clinical practice guidelines[20, 21] and commonly used as a structure for journal and conference abstracts[22, 23, 24]. Additionally, if abstracts meet this defacto standard, it will provide a corpus guarantee for the performance of move recognition.

2.4 Problem Analysis of PubMed 200k RCT

PubMed 200k RCT[12] is consist of 195,654 structured abstracts, which are constructed upon MEDLINE/PubMed Baseline Database published in 2016 based on the two following criteria: (1) belong to RCTs: the abstract must belong to an RCT. This corpus builder relies on the article's MeSH terms only to select RCTs. Specifically, only the articles with the MeSH term D016449, which corresponds to an RCT, are included in PubMed 200k RCT; (2) structured: the abstract must contain between 3 and 9 sections (inclusive), and it should not contain any section labeled as "None", "Unassigned", or "" (empty string). The label of each section was originally given by the authors of the articles, typically following the guidelines given by journals: as many labels exist, PubMed 200k RCT maps them into a smaller set of standardized labels: background, objective, methods, results, conclusions, "None", "Unassigned", or "" (empty string); The original structured abstract format in MEDLINE/PubMed Baseline Database is shown in **Listing 1** below. For each abstract, PubMed 200k RCT detect sentence and token boundaries using the Stanford CoreNLP toolkit[25].

Listing 1. Example of a MEDLINE/PubMed structured abstract.

<Abstract>
<AbstractText Label="Introduction" NlmCategory="BACKGROUND">Asthma is a poorly understood disease. Risk factors are not established, and the natural history of the disease is unknown.</AbstractText>
<AbstractText Label="Aims" NlmCategory="OBJECTIVE">Using subjects of a community-based study, we have prospectively compared young adults destined to develop asthma with control subjects to determine difference…</AbstractText>
<AbstractText Label="Design, setting and participants" NlmCategory="METHODS">Subjects were participants of the Tucson Epidemiologic Study of Airways Obstructive Disease. They were studied between the ages of 15…</AbstractText>
<AbstractText Label="Results" NlmCategory="RESULTS"> Logistic regression showed that "wheeze" and "attacks of shortness of breath with wheeze" were independently predictive of asthma. Positive allergy skin test results…</AbstractText>
<AbstractText Label="Discussion" NlmCategory="CONCLUSIONS"> Logistic regression showed that "wheeze" and "attacks of shortness of breath with wheeze" were independently predictive of asthma. Positive allergy skin…</AbstractText>
</Abstract>

Source: https://www.ncbi.nlm. nih.gov/pubmed/7963152.

In our previous research work, we conducted a series of studies based on PubMed 200K RCT and identified several problems of this dataset. We have summarized these problems into four aspects, which are discussed in Sections 2.4.1 to 2.4.4.

2.4.1 Inappropriate mapping between sentences and NlmCategories

As shown in Table 2, there are some sentence with inappropriate NlmCategory in PubMed 200k RCT. This problem occurs especially among the NlmCategories of "BACKGROUND", "OBJECTIVE" and "METHODS", which has negative effect for the move recognition accuracy and performance. We analyze the possible causes of this problem may be that the author did not give the proper label for their abstract, or caused by the uniformly mapping by NLM-assigned categories. As Listing 1 shows, each structured abstract has two kinds of labels: (1) Label: annotated and provided by the abstracts' authors or journal. In Listing 1, Labels are <Introduction, Aims, Design, setting and participants, Results, Discussion>; (2) NlmCategory: predefined and provided by the U.S. National Library of Medicine (NLM). As shown in Listing 1, NlmCategories include <BACKGROUND, OBJECTIVE, METHODS, RESULTS, CONCLUSIONS>, which mapped from Labels based on the mapping (Structured-Abstracts-Labels-102615.txt[①]) provided by NLM, in which 3032 various kinds of 'Labels' are mapped into 6 'NlmCategory'. Consequently, there might be instances where the Label and NLMCategory have incorrect mapping relationships.

[①] https://lhncbc.nlm.nih.gov/ii/areas/structured-abstracts/downloads/Structured-Abstracts-Labels-102615.txt

Table 2. Sentence is not consistent with their NlmCategory.

PMID	Sentence with inappropriate label	Label	NlmCategory	Right label
21167340	**BACKGROUND** the objective of the study was to determine whether the effects of infarct-related artery (IRA) infusion… dependent on the dose (quantity and mobility) of the cells infused.	BACKGROUND, METHODS, RESULTS, CONCLUSIONS	BACKGROUND, METHODS, RESULTS, CONCLUSIONS	**OBJECTIVE**
19922626	**BACKGROUND** Therefore, the objective of the study was to examine the effect of GERD on HRQOL in adolescents.	BACKGROUND, METHODS, RESULTS, CONCLUSIONS	BACKGROUND, METHODS, RESULTS, CONCLUSIONS	**OBJECTIVE**
24245491	**METHODS** We aim to assess the effectiveness using a randomized design in an outpatient hospital setting.	BACKGROUND, METHODS, DISCUSSION, TRIAL REGISTRATION	BACKGROUND, METHOD, CONCLUSIONS, BACKGROUND	**OBJECTIVE**

We conduct a quantitative analysis of sentences in the PubMed 200k RCT by inviting an annotator, named Annotator0, to perform manual labeling. In the first step, we randomly extract 3,000 abstracts from PubMed 200k RCT. In the second step, Annotator0 performs manual labeling for these extracted abstracts. After manual labeling, the statistics of incorrect abstracts and sentences are shown in Table 3. There are 1,344 abstracts with problematic sentences, accounting for 44.80% of the total. In particular, there are 878 sentences(31.05%) labeled as BACKGROUND that were incorrect, and 884 sentences(30.94%) labeled as OBJECTIVE that were incorrect, which are approximately one-third of the total.

Table 3. Statistics of incorrect abstracts and sentences.

	Error	Original	Percentage
# Abstracts	1344	3000	44.80%
# Sentences	2,234	34,168	6.54%
BACKGROUND	878	2828	31.05%
OBJECTIVE	884	2857	30.94%
METHODS	141	11047	1.28%
RESULTS	232	12082	1.92%
CONCLUSIONS	99	5354	1.84%

2.4.2 Unbalanced distribution of move labels, especially 'Background' and 'Objective'

Figure 1 shows the distribution of move labels in PubMed 200k RCT. We can observe that the number of 'Background' and 'Objective' sentences is noticeably smaller than the other three label sentences. Specifically, these two categories of sentences account for 8.89% and 8.43% respectively, which results in an unbalanced corpus.

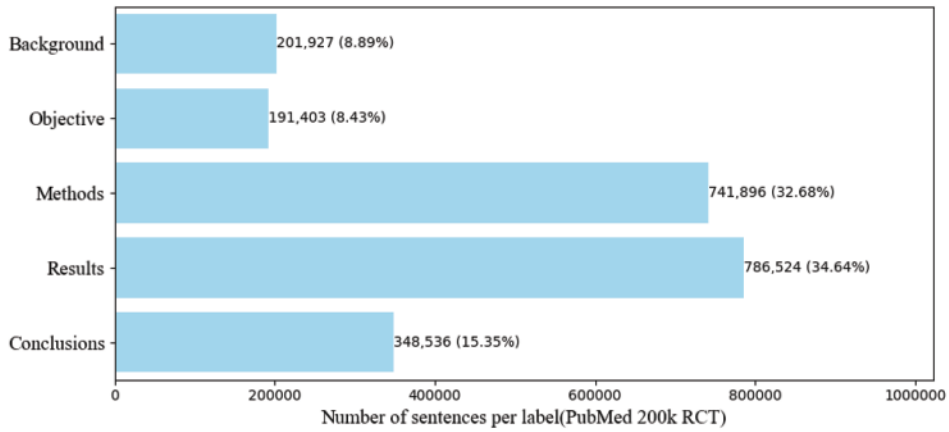

Figure 1. Distribution of move labels in PubMed 200k RCT

In order to analyze the causes of the aforementioned unbalanced problem, we conduct a statistical analysis of the number of abstracts in the PubMed 200k RCT dataset that do not contain specific labels. Table 4 displays the statistics of abstracts without specified labels in PubMed 200k RCT. We can observe that more than half (54.04%) of the abstracts do not have BACKGROUND sentences, and 30.42% of the abstracts do not have OBJECTIVE sentences. We suppose that this causes imbalance in the PubMed 200k RCT dataset leads to move classifiers using PubMed 200k RCT not being able to learn enough features of 'Background' and 'Objective' sentences.

Table 4. Abstracts without specified Labels in PubMed 200k RCT.

Without Labels	#Abstracts	Percentage
BACKGROUND	103021	54.04%
OBJECTIVE	57991	30.42%
METHODS	6608	3.47%
RESULTS	5260	2.76%
CONCLUSIONS	11	0.006%

As reported in the original paper of PubMed 200k RCT[12], based on bi-ANN model, the F1 scores of the five categories are [Background=75.6%, Objective=70.7%, Methods=96.0%, Results=95.2%, Conclusions=94.2%], From these results, we can observe that the performance of Background and Objective is significantly lower than the other three categories. To some extent, we can attribute this situation to the unbalanced distribution of move labels, especially, the lack of sufficient Background and Objective content in the corpus to guide the model to learn useful features about these two Labels and their corresponding sentences. If we construct a corpus that ensures each abstract contains all five move categories, especially requiring the presence of BACKGROUND and OBJECTIVE. we might significantly improve the quality of our corpus.

2.4.3 Experimental registration, ethical and fund information mixed in background

According to Andrade et al.,[26] research, they provide detailed suggestions for writing a good abstract and point out that the laboratory and safety assessments are routinely performed in clinical studies, so that it is unnecessary content in an abstract, unless there is a specific need to highlight these in the abstract. Trial info always occurred in the end of abstract, labeled as background. It contributes little to the abstracts move recognition, but only let the move recognition model be confused to learn the main information typical feature about the 'background' category, in our refined corpus, these trail abstracts should be ignored as much as possible, which might make the refined corpus smaller and finer.

Table 5 shows In PubMed 200k RCT, according to our statistics, there is totally 11,063 abstracts include trial registration, ethics, fund support, etc. account for 5.80% in PubMed 200k RCT trainset, which is totally of 190654 abstracts.

Table 5. Hybridity of registration information, ethics and fund support information.

PMID	Sentence with registration information	Label	NlmCategory
24245491	BACKGROUND ClinicalTrials.gov Identifier : NCT@.	TRIAL REGISTRATION	BACKGROUND
24384905	BACKGROUND Ethical approval was received from the Medical Ethical Committee of the Catharina Hospital Eindhoven, the Netherlands (NL@ @).	ETHICS AND DISSEMINATION	BACKGROUND
24703047	BACKGROUND GlaxoSmithKline	Funding	BACKGROUND
23514036	BACKGROUND ClinicalTrials.gov : NCT@.	Trial registration	BACKGROUND
26116485	BACKGROUND Efficacy and Mechanism Evaluation Programme, funded by the Medical Research Council (MRC) and managed by the National Institute for Health Research (NIHR) on behalf of the MRC-NIHR partnership.	Funding	BACKGROUND
20799286	BACKGROUND UMIN@ (http://www.umin.ac.jp/ctr/index.htm).	Registration number	BACKGROUND
22364685	BACKGROUND Gentium SpA, European Group for Blood and Marrow Transplantation.	Funding	BACKGROUND
24365174	BACKGROUND ClinicalTrials.gov (no.NCT@).	Trial registration	BACKGROUND
23025261	BACKGROUND This trial is registered with Current Controlled Trials and is traceable as ISRCTN@.	Trial registration	BACKGROUND
22152147	BACKGROUND ISRCTN : ISRCTN@.	Trial registration number	BACKGROUND

2.4.4 Uninformative ultra-short sentences

Table 6 shows In PubMed 200K RCT, there are some short/meaningless sentences, which appears in the context of some special characters (such as "+ / -") or abbreviations (such as approx. A.C. chem.), these small text fragments were apparently come from the same one sentence but were split into different sentences, as shown in Table 4. This affects the quality of the corpus to some certain extent. Essentially, these problems is largely due to the methods of sentence split, and the error of sentence segmentation tool kit.

Table 6. Sentence segmentation error.

PMID	Structured Abstracts from PubMed 200k RCT	Original Abstracts
24444257	RESULTS An increased risk (risk ratio > @) to occur at large abscess lesions was observed for Prevotella (P.) oralis, P. buccae, P. oris, P. intermedia, Fusobacterium nucleatum and Streptococcus (Strep.) RESULTS anginosus group. RESULTS An increased risk to occur at large infiltrate lesions was found for Strep. RESULTS salivarius, Strep. RESULTS parasanguis, Strep. RESULTS anginosus group, Capnocytophaga spp., Neisseria (N.) sicca, Neisseria spp., Staphylococcus (Staph.) RESULTS aureus, P. intermedia, P. buccae, Prevotella spp. RESULTS and P. melaninogenica.	An increased risk (risk ratio >1) to occur at large abscess lesions was observed for Prevotella (P.) oralis, P. buccae, P. oris, P. intermedia, Fusobacterium nucleatum and Streptococcus (Strep.) anginosus group. An increased risk to occur at large infiltrate lesions was found for Strep. salivarius, Strep. parasanguis, Strep. anginosus group, Capnocytophaga spp., Neisseria (N.) sicca, Neisseria spp., Staphylococcus (Staph.) aureus, P. intermedia, P. buccae, Prevotella spp. and P. melaninogenica.
9688247	METHODS Local estrogen, physiotherapy and electrostimulation combined with close follow-up. METHODS @. METHODS Change in severity of incontinence from start of treatment (index range @-@). METHODS @. METHODS Change in impact from start of treatment (index range @-@). METHODS @. METHODS Quantitative measures in relation to micturition. METHODS @. METHODS Criteria based classification into cured, improved, unchanged, worse.	Intervention: Local estrogen, physiotherapy and electrostimulation combined with close follow-up. 1. Change in severity of incontinence from start of treatment (index range 0-8). 2. Change in impact from start of treatment (index range 0-4). 3. Quantitative measures in relation to micturition. 4. Criteria based classification into cured, improved, unchanged, worse.
2183665	METHODS Mean glomerular filtration rate + / - standard error was @ + / - @ mL/s. METHODS m@ (@ + / - @ mL/min @ m@) ; mean effective renal plasma flow was @ + / - @ mL/s. METHODS m@ (@ + / - @ mL/min @ m@).	Mean glomerular filtration rate +/- standard error was 0.36 +/- 0.03 mL/s.m2 (37 +/- 3 mL/min.1.73 m2); mean effective renal plasma flow was 1.6 +/- 0.18 mL/s.m2 (166 +/- 19 mL/min.1.73 m2).
10716474	METHODS and T (L-arg. METHODS + T) or an inactive control group (C).	Forty patients with severe CHF (left ventricular ejection fraction 19 +/- 9%) were randomized to an L-arg. group (8 g/day), a training group (T) with daily handgrip training, L-arg. and T (L-arg. + T) or an inactive control group (C).

2.5 Summary

With the above analysis, we consider that the (1)(2) problems above is mainly because that the structured abstracts in PubMed 200k RCT doesn't constraint the number of NLMCategory. If each abstract has five

NLMCategory(five-move scheme): BACKGROUND, OBJECTIVE, METHODS, RESULTS, CONCLUSIONS, these problems will be reduce. The (3) problem often appeared in the abstracts where the NLMCategory order is not as 'B-O-M-R-C', so if we control the NLMCategory appear order, this problem will be largely solved. As for (4) problem, we can design refining rules to correct the results of sentence segmentation results, combine fragments from the same sentence to reconstitute a complete sentence.

3. CORPUS CONSTRUCTION

We construct our Refined Corpus for Move Recognition in general medical field mainly following two steps below. Firstly, Explain the process of data acquisition and the criteria of data selection; Secondly, describe the data preprocess and sentence segmentation. Thirdly, we give an overview of our refined corpus statistics. At the end of this section, we conduct a Human Reviewing and Consistency Check about out RCMR corpus.

3.1 Structured Abstracts Acquisition and Selection

As a large corpus of annotated abstracts, we built our corpus using structured abstracts extracted from PUBMED/MEDLINE 2020 Baseline (1,015 xml files). In the first acquisition step, we collected 4,445,664 structured abstracts. All collected abstracts(as Listing1 in Section 2.4.1) were written in English and annotated with labels provided by the abstract's authors and NlmCategory mapping by NLM. The kinds of different NlmCategory contained in an abstract is interpreted as its rhetorical structure, and the sequence of NlmCategory observed in an abstract was interpreted as its rhetorical sequence.

Different NlmCategories serve distinct functions within an abstract. Specifically, the Background establishes the research context and situates the current research question within the entire scientific research landscape; the Objective proposes the research question; the Methods details the methodology, including procedures, experimental instruments, and subjects; the Results summarize the experimental findings; and the Conclusions discuss the research implications. These five Categories are all essential components of a scientific research paper abstract. Following the strategy of Abdollahpour et al.[27] and Hirohata et al. [28], we build our refined corpus based on the three assumptions below: (1) An abstract that has five different NlmCategories sections, as shown in Listing 1, with each section clearly distinguished from the others and reducing the possibility of inappropriate Label-NlmCtegory mapping, can effectively inform readers about the main points of the paper. We name it five-NlmCategories scheme, which is usually required by academic writers in most scholarly journals and provides a comprehensive description of research. Conversely, an abstract with fewer than five NlmCategories may negatively impact fine-grained analysis for various downstream text-mining tasks, such as information extraction, information retrieval, and automatic summarization; (2) An abstract containing five NlmCategories, with each NlmCategory appearing in a linear fashion, i.e., BACKGROUND-OBJECTIVE-METHODS-RESULTS-CONCLUSIONS, constitutes a well-structured and standard format reflecting the scientific discovery process. Such a structure ensures essential guarantee for the performance of move recognition task; (3) If our refined corpus is

selected for use in the medical subdomain, such as PubMed 200k RCT, it will provide data support and expand the available corpus for conducting further fine-grained text mining and exploration in the medical subfield. This adaptability can vertically increase research depth and precision in move recognition.

Based on aforementioned assumptions and considerations, we construct our refined move recognition corpus, RCMR, with three criteria:

(1) Structure Rule: We utilize five-move scheme, each abstract must contain and only contain five NlmCategories sections, that is: BACKGROUND(B), OBJECTIVE(O), METHODS(M), RESULTS(R), CONCLUSIONS(C);

(2) Sequence Rule: Each abstract sections must be ordered by 'B->O->M->R->C'. Specifically, we use SQL Server 2012 to manage the acquired structured abstracts. Firstly, we join the different NLMCategories of each abstract with "||" symbol, then filter the structured abstracts using SQL statement, as Figure 2 shows, to select structured abstracts containing BACKGROUND(B), OBJECTIVE(O), METHODS(M), RESULTS(R), CONCLUSIONS(C), and NlmCategory appearing in a linear fashion;

```
SELECT DISTINCT [MedlineCitation_PMID]
FROM [PM2020_Structured_abstracts].[dbo].[PMID_ConcatLabels]
WHERE [NlmCategory]='BACKGROUND||OBJECTIVE||METHODS||RESULTS||CONCLUSIONS'
```

Figure 2. SQLQuery to select refined structured abstracts

(3) RCT Rule(for RCMR_RCT): we construct a subset of RCMR, named RCMR_RCT, each abstract in this subset must belong to randomized controlled trials (RCT, MeSH term: D016449).

3.2 Sentence Segmentation Processing

After above filtering process, our corpus was reduced to 28,8436 abstracts. The resulting corpus was segmented into sentences by using Python package ScispaCy, which is containing spaCy models for processing biomedical, scientific or clinical text. In view of the problems of PubMed 200k RCT analysis above, we design five sentence splitting rules below, to solve the incorrect sentence segmentation problems, which is based on the results of ScispaCy[29] to correct the errors appear in the sentence splitting of English scientific literature abstracts. Finally, we get totaling 3,386,008 sentences, in which B:561480, O:313525, M:823990, R:1143588, C:543425.

1) If the sentence is beginning with Arabic numerals, such as '1', '2' and so on, it will be merged into the next sentence.
2) If the sentence ends with a "+/-" symbol, it will be merged into the next sentence.
3) If the sentence starts with a lowercase letter, it is merged into the previous sentence.
4) If the sentence has less than 30 characters, it is merged into the previous sentence.
5) If the sentence contains ")" and the number of '(' and ')' in the previous sentence does not add up to an even number, it will be merged into the previous sentence.

After the above sentence segmentation processing, the corpus is split into three text files: one for the trainset, one for the test-set, and one for the devset, as shown in Table 7. Each file has the same format: Each sentence was annotated with a NlmCategory corresponding to the NlmCategory of the section to which the sentence belongs, i.e., all sentences in Methods sections were labeled as Methods, and so forth. And Each line corresponds to either a PMID or a sentence with its capitalized NLMCategory at the beginning. Each sentence and its NLMCategory is separated by a space, and replace numbers with "@".

Table 7. Refined Corpus Discription.

Subset	# NlmCategory	# Abstracts	# Sentences
trainset	5	283,436	3,327,126
testset	5	2,500	29,375
devset	5	2,500	29,494
Total	5	28,8436	3,386,008

3.3 Refined Corpus Statistics

After above data acquisition, selection and data preprocessing, we finally acquired our refined corpus, named RCMR(Refined Corpus for Move Recognition Based on PubMed). Figure 3 shows an example of abstract from our corpus along with its sentences' labels. Each abstract has the same format: each line corresponds to either a PMID or a sentence with its capitalized label at the beginning. Label and sentence are separated by a '\t' space.

```
###1552048
BACKGROUND  The safety of psoralen plus ultraviolet A (PUVA) light therapy has been an issue of debate.
BACKGROUND  A few multiple-center cooperative studies have reported an increase of basal cell and squamous cell carcinomas among
BACKGROUND  In our institute, more than 1000 patients have been treated with PUVA since 1975.
OBJECTIVE   We investigated the incidence of skin cancer among patients who received high doses of PUVA to see whether such incid
METHODS This is a historical cohort study of two comparison groups of patients.
METHODS Subjects under study were 492 psoriasis patients who received PUVA treatments between 1975 and 1989.
METHODS One group of 103 patients, defined as the high-dose group, received an accumulated PUVA dose of 1000 joules/cm2 or more;
METHODS The occurrence of skin cancer in the two comparison groups is analyzed.
RESULTS In the high-dose group we observed an increased number of patients with squamous cell carcinoma, keratoacanthoma, and acti
RESULTS We did not see any patients with genital cancer, melanoma, or an increased number of patients with basal cell carcinoma.
CONCLUSIONS The risk of squamous cell carcinoma developing in patients who received a high dose of PUVA is confirmed.
CONCLUSIONS We speculate a combination of factors, including PUVA, may contribute to this risk.

###7514436
BACKGROUND  The role of interleukin-3 (IL-3) in stimulating the growth of early myeloid progenitor cells is very well established.
BACKGROUND  Therefore, IL-3 has been incorporated into many post-bone-marrow transplantation and intensive chemotherapy programs
BACKGROUND  However, the effect of IL-3 on normal and malignant lymphocytes has not been well studied.
OBJECTIVE   The purpose of this study was to evaluate the in vitro effect of IL-3 on the growth of follicular small-cleaved-cell
METHODS IL-3 receptor expression on the surface of CD19+ cells was determined by two-color flow cytometry measuring the receptor-
METHODS Seven cases of FSCCL were compared to six normal controls.
METHODS Cell proliferation was evaluated by [3H]thymidine incorporation into cells grown in suspension cultures.
RESULTS All seven cases of FSCCL expressed the IL-3 receptor on the surface of CD19+ cells, whereas all six cases of CD19+ cells
RESULTS IL-3 had antiproliferative activity against FSCCL as manifested by a decrease in [3H]thymidine incorporation and a decrea
CONCLUSIONS IL-3 inhibits the growth of FSCCL cells in vitro.
CONCLUSIONS Clinical trials to evaluate the in vivo effect of IL-3 in patients with FSCCL are warranted.
```

Figure 3. Our refined corpus example (PMID: 1552048 and 7514436).

In order to have a more comprehensive and clear understanding of our RCMR corpus, we plot statistic figures below, Figure 4 shows the distribution of the number of sentences per abstract. Figure 5 shows the distribution of the number of tokens the sentence.

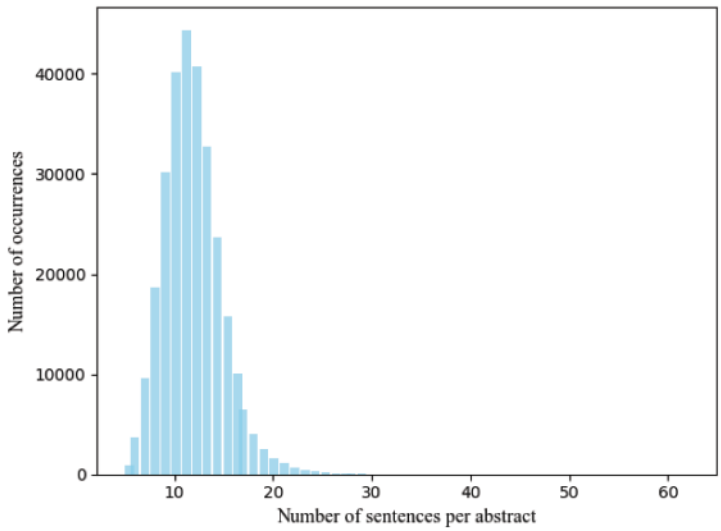

Figure 4. Distribution of the number of sentences per abstract. Minimum:5.00; mean:11.74; maximum:62.00; variance:8.93; skewness:1.04; kurtosis:6.74.

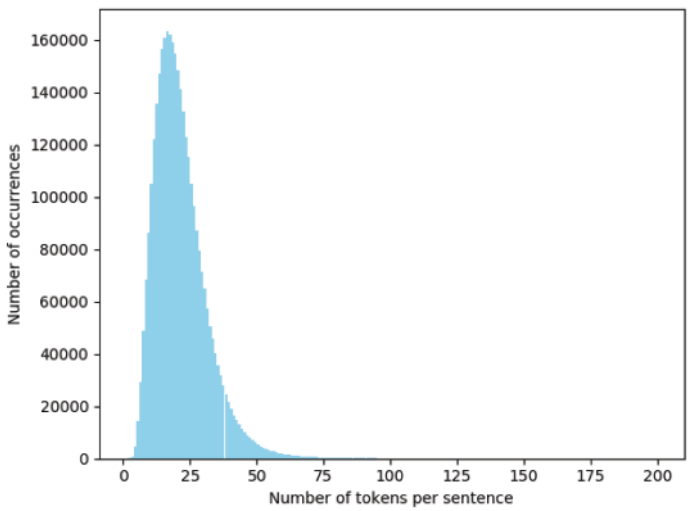

Figure 5. Distribution of the number of tokens the sentence. Minimum: 1.00; mean: 21.11; maximum: 335.00; variance: 202.80; skewness: 1.68; kurtosis: 10.68.

Figure 6 provides a count of the number of sentences per label within our RCMR corpus. We observe that the most prevalent label (Results) accounts for 33.77% of the sentences, while the least frequent label (Objective) accounts for 9.26%, which indicates that our RCMR is still unbalanced, although it is better than PubMed 380k's and PubMed 200k RCT's. As Table 8 shows, In PubMed 380k, The number of 'Results' and 'Objective' sentences account for 32.75% and 8.69%, respectively. While in PubMed 200k RCT, these numbers are 34.64% and 8.43%, respectively.

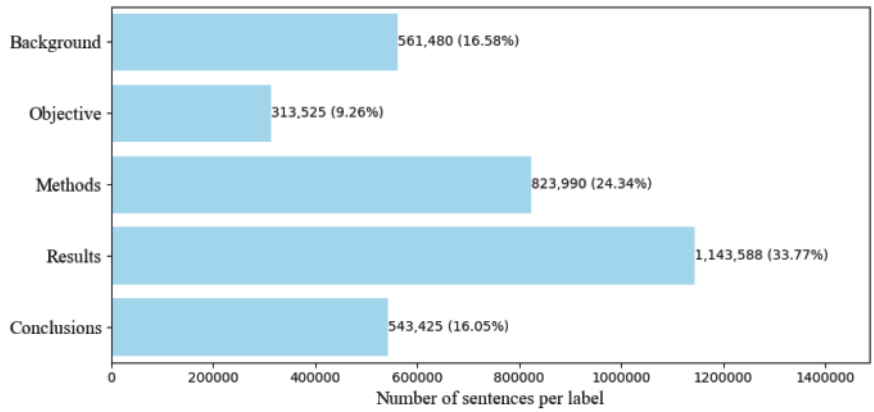

Figure 6. Number of sentences per label of our RCMR.

Table 8. Detailed Category distribution of our RCMR, PubMed 380k, and PubMed 200k RCT.

NlmCategory	RCMR	PubMed 380k	PubMed 200k RCT
# Background (%)	561,480(16.58%)	770,009(15.64%)	201,927(8.89%)
# Objective (%)	313,525(9.26%)	427,928(8.69%)	191,403(8.43%)
# Methods (%)	823,990(24.34%)	1,306,519(26.53%)	741,896(32.68%)
# Results (%)	1,143,588(33.77%)	1,612,851(32.75%)	786,524(34.64%)
# Conclusion (%)	543,425(16.05%)	806,730(16.38%)	348,536(15.35%)
# Total(%)	3,386,008(100%)	4,924,037(100%)	2,270,286(100%)

This "unbalanced distribution" problem is expected, since authors typically tend to emphasize their findings in their writing, resulting in more extensive Results sections. On the other hand, they tend to be more concise when describing their study's purpose, which lead to Objective sections containing only a few sentences.

3.4 Human Review and Consistency Evaluation

To ensure the reliability and accuracy of our RCMR corpus, it is essential to conduct a human review and consistency evaluation. We invite two independent annotators to manually review the corpus and verify the mapping between the sentence content and their respective NlmCategory labels. By assessing the agreement between the annotators' labels and the original NlmCategory labels, we can determine the consistency and overall quality of our RCMR corpus. Any discrepancies or inaccuracies identified during this evaluation process should be rectified. Specifically, The specific annotation steps are as follows:

1) Randomly select 2,000 structured abstracts(21861 sentences) from our RCMR;
2) Designate the correspondence between NlmCategory and numerical labels as follows: Background=1, Objective=2, Method=3, Results=4, Conclusion=5. As shown in Figure 7, the green column is the column that needs to be checked. If the sentence content does not match the label content, it should

be modified to the corresponding number. For example, if a sentence expressing the meaning of "research purpose" is incorrectly marked as 1, it needs to be manually modified to 2. The red column, the blue column, and last column do not require any modification and are for reference only.

3) Invite two PhD students (e.g., Annotator1 and Annotator2) to manually annotate the selected corpus according to the above annotation explain and criteria.
4) Calculate consistency indicators and compare the results between human annotation and the existing NLM Category system.

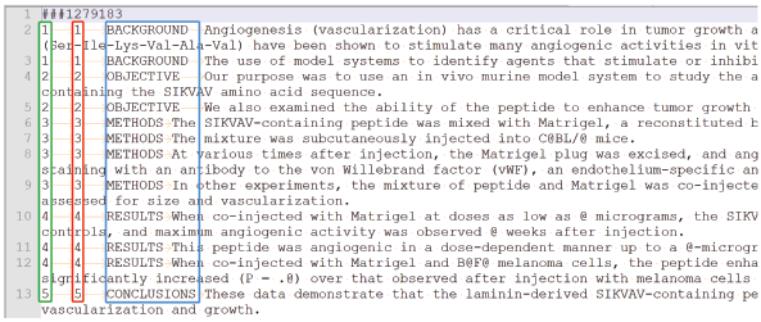

Figure 7. Annotator amended examples.

By following above four steps, the quality of our RCMR can be assessed, and any discrepancies or inaccuracies can be identified and rectified.

Table 9 calculate and compare the kappa[30] coefficients between PubMed 200k RCT (mentioned in Section 2.4.1 and annotated by Annotator 0) and our RCMR annotations, which are conducted by Annotator 1 and Annotator 2.

Table 9. Kappa coefficients between PubMed 200k RCT and our RCMR.

Reviewer	Corpus	Kappa
Annotator 0 (Section 2.4.1)	PubMed 200K RCT	0.9117
Annotator 1	RCMR (ours)	0.9955
Annotator 2	RCMR (ours)	0.9959

Table 10 and Table 11 present the confusion matrix results of manual annotations conducted by Annotator1 and Annotator2. We observe that the Kappa coefficients, which evaluate the agreement between the manual labeling and our RCMR corpus, exceed 99%, significantly surpassing those of the PubMed 200K RCT. This observation suggests that there are fewer discrepancies between the two independent annotators, and our RCMR corpus demonstrates a strong mapping correlation between NlmCategories and their corresponding sentence content. Consequently, this indicates that our RCMR corpus is of high quality and exhibits adaptability for move recognition tasks.

Table 10. Results of Human Reviewing (1) between annotator and RCMR.

RCMR	Annotator 1					total
	B	O	M	R	C	
B	3643	0	0	0	0	3643
O	3	2167	2	0	0	2172
M	0	0	4896	9	0	4905
R	1	0	11	7580	38	7630
C	0	0	0	12	3499	3511
total	3647	2167	4909	7601	3537	21861
Kappa	0.9955					

Note: B: BACKGROUND, O: OBJECTIVE, M: METHODS, R: RESULTS, and O: CONCLUSIONS.

Table 11. Results of Human Reviewing (2) between annotator and RCMR.

RCMR	Annotator 2					total
	B	O	M	R	C	
B	3639	4	0	0	0	3643
O	9	2155	8	0	0	2172
M	0	4	4898	3	0	4905
R	0	0	14	7615	1	7630
C	0	0	0	26	3485	3511
total	3648	2163	4920	7644	3486	21861
Kappa	0.9959					

Note: B: BACKGROUND, O: OBJECTIVE, M: METHODS, R: RESULTS, and O: CONCLUSIONS.

4. CORPUS PERFORMANCE

4.1 Experimental Design and Datasets

The aim of our experiments is to evaluate the quality of our refined corpora, RCMR and RCMR_RCT. We design experiments from three distinct perspectives, corresponding to three experimental groups, as detailed in Table 12. Specifically:

Exp1: Assess the overall performance of our RCMR using various sentence classification models. We utilize the RCMR trainset (D1), which comprises 283,436 refined structured abstracts from the entire medical domain. Four popular sentence classification models are trained using this dataset. Subsequently, we evaluate the classification performance of these trained models on the RCMR testset (T1), which consists of 29,375 structured sentences.

Exp2: Compare our RCMR with PubMed 380k, as both corpora encompass the entire medical domain. Based on the analysis in Section 2, the data volumes of RCMR and PubMed 380k are inconsistent. To ensure

comparability of experimental results, we randomly sample a subset from the RCMR trainset (D2) containing 15,000 refined structured abstracts from the entire medical domain. Similarly, we randomly sample another subset from PubMed 380k (D3) comprising 15,000 structured abstracts from the entire medical domain. We train BERT and MSMBERT models using D2 and D3, respectively. The classification performance of the trained models is then tested on the RCMR testset (T1), as in Exp1.

Exp3: Compare our RCMR_RCT with PubMed 200k RCT, as both corpora are from the medical subdomain of randomized controlled trials (RCTs). Similar to Exp2, there are differences in data volume between RCMR_RCT and PubMed 200k RCT. To make experimental results comparable, we randomly sample a subset from the RCMR_RCT trainset (D4) consisting of 15,000 refined structured abstracts in the medical subdomain of RCTs. We then randomly sample a subset from PubMed 200k RCT (D5), containing 15,000 structured abstracts in the medical subdomain of RCTs. BERT and MSMBERT models are trained using D4 and D5, respectively. The classification performance of the trained models is tested on the RCMR_RCT testset (T2), which includes 29,566 structured sentences.

Table 12. Experiment No. and Experimental Dataset Discriptions.

Exp No.	Data No.	Trainset Description	Testset Description # (Testset Sentences)
Exp1	D1	our RCMR trainset, entire medical domain. All in use.	RCMR testset (29,375)
Exp2	D2	our RCMR trainset, entire medical domain. Random sampling 15k abstracts.	RCMR testset (29,375)
	D3	PubMed 380k, entire medical domain. Random sampling 15k abstracts.	
Exp3	D4	our RCMR_RCT trainset, medical subdomain of RCTs. Random sampling 15k abstracts.	RCMR_RCT testset (29,566)
	D5	PubMed 200k RCT trainset, medical subdomain of RCTs. Random sampling 15k abstracts.	

4.2 Comparison Method

Attention-BiLSTM: Attention-BiLSTM[7] refers to the Bidirectional Long Short-Term Memory Network based on the Attention Mechanism. We adopt the method from our previous study [31, 32], which employs a five-layer Attention-BiLSTM model, This model can capture text context features from both previous and subsequent directions, allowing it to automatically learn feature words in each sentence

HSLN: We use the HSLN-RNN model presented by Di Jin et al. [9] from MIT. This model is a hierarchical sequential labeling network named HSLN-RNN, which leverages the contextual information from surrounding sentences to classify the current sentence. It employs a Bi-LSTM layer after encoding sentence-level features to capture contextual features within sentences and a CRF layer to capture sequential features within surrounding move labels.

BERT: BERT (Bidirectional Encoder Representations from Transformers)[10] was released by Google in October 2018 and has since received widespread attention due to its record-breaking performance in 11 NLP tasks upon released. We use the BERT-base model, which has a hidden size of 768, 12 Transformer blocks[33], and 12 self-attention heads. We fine-tune BERT with the following settings: a batch size of 5, a max sequence length of 512, the learning rate of 3e-5, the init_checkpoint of BERT_base, 100,000 training steps, and 10,000 warm-up steps.

MSMBERT: MSMBERT (Masked Sentence Model based on BERT) is presented in our previous work[17]. This model effectively captures both content features and contextual features of sentences. It is easy to implement, as it only necessitates reconfiguring the input data without altering the structure of neural networks. Furthermore, it utilizes the same BERT-base model and hyperparameter setting as described above.

4.3 Evaluation Metrics

We use evaluation metrics, namely precision (P), recall (R), and F1 score in each experiment, and the results and analysis of experiments Exp1 to Exp3 are discussed in Section 4.4.

4.4 Results

The Results of Exp1: Assess the overall performance of our RCMR using various sentence classification models

Table 13 presents the results of Exp1, which evaluates the overall performance of our RCMR using various sentence classification models, as mentioned in Section 4.2. We can observe that all these sentence

Table 13. Overall performance of RCMR using various sentence classification models.

		B	O	M	R	C	Macro avg
Attention Bi-LSTM (D1)	P	0.9378	0.8895	0.9291	0.8281	0.7832	0.8735
	R	0.9136	0.8997	0.9229	0.8704	0.7423	0.8698
	F1	0.9256	0.8945	0.9260	0.8487	0.7622	0.8714
HSLN (D1)	P	0.9912	0.9901	0.9814	0.9595	0.9820	0.9808
	R	0.9967	0.9649	0.9782	0.9834	0.9437	0.9734
	F1	0.9940	0.9774	0.9798	0.9713	0.9625	0.9770
BERT (D1)	P	0.8416	0.9518	0.9455	0.9040	0.8258	0.8937
	R	0.9156	0.9180	0.9335	0.9120	0.7679	0.8894
	F1	0.8770	0.9346	0.9394	0.9080	0.7958	0.8910
MSMBERT (D1)	P	0.9728	0.9785	0.9648	0.9541	0.9625	0.9634
	R	0.9920	0.9564	0.9739	0.9640	0.9203	0.9634
	F1	0.9823	0.9673	0.9693	0.9590	0.9409	0.9633
	support	4869	2708	7170	9937	4691	29375 (T1)

Note: (D1): our RCMR, (T1): our RCMR testset. B: BACKGROUND, O: OBJECTIVE, M: METHODS, R: RESULTS, and O: CONCLUSIONS.

classification models perform quite well, with F1 scores exceeding 87%: 87.14%, 97.70%, 89.10%, and 96.33%, respectively. It is also noteworthy that the recognition results for BACKGROUND and OBJECTIVE sentences are comparable to those of METHODS, RESULTS, and CONCLUSIONS. Taking the MSMBERT model and its recognition results as an example, the recognition accuracy for BACKGROUND and OBJECTIVE sentences is [97.28%, 97.85%], and for METHODS, RESULTS, and CONCLUSIONS, it is [96.48%, 95.41%, 96.25%]. Both the former and the latter achieve above 95% high accuracy, with insignificant differences.

Considering the analysis of PubMed 200k RCT in Section 2, the recognition accuracy of bi-ANN trained on PubMed 200k RCT for BACKGROUND and OBJECTIVE is [70.7%, 77.1%], while for METHODS, RESULTS, and CONCLUSIONS it can reach [95.5%, 95.6%, 94.6%]. This indicates that the recognition accuracy of BACKGROUND and OBJECTIVE is significantly lower than that of other categories, further suggesting that sentence classification models struggle to learn and capture informative features to accurately distinguish BACKGROUND and OBJECTIVE from other categories.

In comparison, our RCMR has the advantage of high-quality and more accurate semantic mapping between sentences and categories, especially for BACKGROUND and OBJECTIVE. Therefore, using RCMR, sentence classification models can achieve better performance with higher recognition accuracy across all categories, which narrows the recognition accuracy gap between different categories and further enhances the performance of sentence classification models.

The Results of Exp2: Compare with PubMed-380k, in the entire medical domain

Table 14 presents the results of Exp2, designed in Section 4.1, where we compare our RCMR(D2) and PubMed 380k(D3) using BERT and MSMBERT models. In this paper, RCMR consists of abstracts from the

Table 14. D2 and D3 performance on BERT and MSMBERT.

Classifier model		B	O	M	R	C	Macro avg
BERT (D2)	P	0.8259	0.9161	0.9081	0.8907	0.7669	0.8668
	R	0.8624	0.9110	0.9245	0.8779	0.7363	0.8672
	F1	0.8438	0.9135	0.9162	0.8843	0.7513	0.8668
BERT (D3)	P	0.8141	0.9134	0.8783	0.9045	0.7517	0.8595
	R	0.8651	0.9114	0.9325	0.8396	0.7472	0.8583
	F1	0.8388	0.9124	0.9046	0.8708	0.7494	0.8582
MSMBERT (D2)	P	0.9653	0.9715	0.9514	0.9448	0.9529	0.9536
	R	0.9885	0.9553	0.9660	0.9498	0.9054	0.9536
	F1	0.9768	0.9633	0.9586	0.9473	0.9285	0.9534
MSMBERT (D3)	P	0.9577	0.9612	0.9323	0.9482	0.9379	0.9454
	R	0.9805	0.9505	0.9724	0.9266	0.9043	0.9454
	F1	0.9805	0.9558	0.9519	0.9373	0.9208	0.9452
	support	4869	2708	7170	9937	4691	29375 (T1)

Note: (D2): our RCMR, (D5): PubMed 380k, (T1): our RCMR testset. B: BACKGROUND, O: OBJECTIVE, M: METHODS, R: RESULTS, and O: CONCLUSIONS.

entire medical domain, making it more general for use. We can observe that BERT and MSMBERT, trained on RCMR, achieve average F1 score improvements compared to those trained on PubMed 380k, by 0.86% and 0.82%, respectively, but the improvement effect is not substantial. Specifically, BERT(D2) achieves a better F1 score than BERT(D3) for BACKGROUND, METHODS, RESULTS categories, with improvements of 0.5%, 1.16%, and 1.35%, respectively. MSMBERT(D2) achieves a better F1 score than MSMBERT(D3) using RCMR results in better F1 score improvements for OBJECTIVE, RESULTS and CONCLUSIONS categories, with improvements of 0.75%, 1%, and 0.82%, respectively. These results somewhat demonstrate that our RCMR is of higher quality than PubMed 380k corpus.

The Results of Exp3: Compare with PubMed 200k RCT, in the medical subdomain of RCTs

Table 15 presents the results of Exp3, designed in Section 4.1, where we compare our RCMR_RCT(D4) and PubMed 200k RCT(D5) using BERT and MSMBERT models. We can observe that BERT(D4) and MSMBERT(D4), trained on RCMR_RCT, achieve better performance than those trained on PubMed 200k RCT. Additionally, as researchers, we know that the research background and objective in an abstract are essential for understanding the main point of a paper. Therefore, it is necessary to improve classification model's recogniton accuracy, especially for these two categories, so that it can offer more useful functions and application value for users. With more attention given to BACKGROUND, OBJECTIVE, BERT(D4) and MSMBERT(D4) improve the recognition accuracy of these two categories.

Table 15. D4 and D5 performance on BERT and MSMBERT.

Classifier model		B	O	M	R	C	macro avg
BERT (D4)	P	0.9182	0.9082	0.9439	0.9247	0.8499	0.9159
	R	0.9105	0.9409	0.9556	0.9274	0.8152	0.9164
	F1	0.9143	0.9243	0.9497	0.9261	0.8322	0.9160
BERT (D5)	P	0.8676	0.8949	0.9517	0.9376	0.8597	0.9147
	R	0.8879	0.8131	0.9631	0.9348	0.8719	0.9147
	F1	0.8776	0.8520	0.9573	0.9362	0.8657	0.9145
MSMBERT (D4)	P	0.9825	0.9754	0.9670	0.9597	0.9701	0.9683
	R	0.9916	0.9595	0.9774	0.9653	0.9403	0.9683
	F1	0.9870	0.9674	0.9722	0.9625	0.9550	0.9683
MSMBERT (D5)	P	0.7676	0.5733	0.9270	0.9455	0.9200	0.8751
	R	0.6499	0.6187	0.9715	0.9557	0.9154	0.8764
	F1	0.7039	0.5952	0.9487	0.9505	0.9177	0.8748
	support	4524	2691	8422	9270	4659	29566 (T2)

Note: (D4): our RCMR_RCT, (D5): PubMed 200k RCT, (T2): our RCMR_RCT testset. B: BACKGROUND, O: OBJECTIVE, M: METHODS, R: RESULTS, and O: CONCLUSIONS.

Specifically, When training the BERT model using RCMR_RCT, the average F1 scores outperform that of PubMed 200k RCT by 0.15%. The BERT model trained on RCMR_RCT achieves a higher F1 score than that of PubMed 200k RCT, especially on BACKGROUND and OBJECTIVE categories, by 3.67% and 7.23%,

respectively; When training the MSMBERT model using RCMR_RCT, the average F1 scores outperform that of PubMed 200k_RCT by 9.35%. The MSMBERT model trained on RCMR_RCT shows better improvement in each category, especially in the BACKGROUND, OBJECTIVE, and CONCLUSIONS categories, where the average F1 scores improves by 28.31%, 37.22%, and 3.73%, respectively. This proves that our RCMR_RCT corpus, which follows sequence role, is more suitable for the MSMBERT model, which considers the contextual features of the sentences.

5. CONCLUSION AND FUTURE WORK

In this paper, we first analyze existing structured abstracts datasets for sentence classification, with a focus on PubMed 200k RCT. We then propose a feasible selection method for obtaining refined PubMed structured abstracts. By utilizing this method, we construct two high-quality refined corpora for move recognition in the entire medical domain and the medical subdomain of randomized controlled trials (RCTs), named RCMR and RCMR_RCT, respectively. Our method, which employs Structure Rules, Sequence Rules, and RCT Rules, ensures the selected PubMed corpus is more standardized and adheres to scientific paper writing norms. In addition, we address specific issues identified in the results of ScispaCy sentence segmentation processing to reduce the occurrence of ultra-short sentences in our refined corpus, ultimately leading to more complete sentence structures and more informative sentences.

To validate the quality of our corpus, we conduct experiments from various perspectives, including overall performance on the full dataset and subset performance in the medical and RCT fields. Our results indicate that our corpus achieves a higher F1 score performance than the PubMed 380k(outperforming by 0.82 points) and PubMed 20k RCT(outperforming by 9.35 points), as evaluated using the MSMBERT model.

Our study has been integrated into the SciAIEngine (http://sciengine.las.ac.cn/)[34], an open platform for researchers to perform move recognition tasks. As future work, we aim to improve the practical application performance of our corpus for unstructured abstract move recognition by gathering and leveraging user feedback. Furthermore, we plan to explore the scalability of our corpus by applying it to other models and tasks.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available in the ScienceDB repository: https://doi.org/10.57760/sciencedb.07524. To reuse the data, please cite the data as: Li, J., et al.: RCMR 280k: Refined Corpus for Move Recognition Based on PubMed Abstracts. Data Intelligence xx(xx), 2023. doi: https://doi.org/10.57760/sciencedb.07524.

ACKNOWLEDGMENTS

This work is supported by the project "Deep learning-based scientific literature knowledge engine demonstration system" (Grant No. E0290905) from the Chinese Academy of Sciences.

AUTHOR CONTRIBUTIONS

Jie Li (lijie201909@mail.las.ac.cn) and Gaihong Yu(yugh@mail.las.ac.cn) designed and produced the whole research and wrote the manuscript. Jie Li performed the research, designed the corpus refinement methodology, conduct data analysis, construct our corpus. Gaihong Yu designed the experiments, carried out the experiment, analysis the results. Zhixiong Zhang(zhangzhx@mail.las.ac.cn) proposed the research problems, supervised the research and provided insightful revision on the manuscript. All authors done a lot of modifications and improvements and finally completed the paper.

REFERENCE

[1] Teufel, S., et al. : An annotation scheme for discourse-level argumentation in research articles. In: Proceedings of The Ninth Conference on European Chapter of the Association for Computational Linguistics, pp. 110–117 (1999)

[2] Hirohata, K., et al.: Identifying sections in scientific abstracts using conditional random fields. In: Proceedings of the Third International Joint Conference on Natural Language Processing, pp. 381–388 (2008)

[3] Yamamoto, Y., et al.: A sentence classification system for multi-document summarization in the biomedical domain. In: Proceedings of International Workshop on Biomedical Data Engineering, pp. 90–95 (2005).

[4] Ding, L. P., et al.: Research on Factors Affecting the SVM Model Performance on Move Recognition. Data Analysis and Knowledge Discovery 3(11), 16–23 (2019)

[5] Fisas, B., et al.: A multi-layered annotated corpus of scientific papers. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), pp. 3081–3088 (2016)

[6] Hochreiter, S., et al.: Long Short-Term Memory. Neural Compuation 9(8), 1735–1780 (1997)

[7] Zhou, P., et al.: Attention-Based Bidirectional Long Short-Term Memory Networks for Relation Classification. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pp. 207–212 (2016).

[8] Kim, Y.: Convolutional Neural Networks for Sentence Classification. arXiv preprint arXiv 14085882 (2014)

[9] Jin, D., et al.: Hierarchical Neural Networks for Sequential Sentence Classification in Medical Scientific Abstracts. arXiv preprint arXiv 180806161 (2018)

[10] Devlin, J., et al.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv preprint arXiv 181004805 (2018)

[11] Kim, S. N., et al.: Automatic classification of sentences to support evidence-based medicine. BMC Bioinformatics 12(S2), S5 (2011)

[12] Dernoncourt, F., et al.: PubMed 200k RCT: a Dataset for Sequential Sentence Classification in Medical Abstracts. In: Proceedings of the The 8th International Joint Conference on Natural Language Processing, IJCNLP, pp. 308–313 (2017)

[13] Ammar, W., et al.: Construction of the literature graph in semantic scholar. In: Proceedings ofthe Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), pp. 84–91 (2018)

[14] Moura, G. B., et al.: Using LSTM Encoder-Decoder for Rhetorical Structure Prediction. In: 2018 7th Brazilian Conference on Intelligent Systems (BRACIS), pp. 278–283 (2018)

[15] Stead, C., et al.: Emerald 110k: A multidisciplinary dataset for abstract sentence classification. In: Proceedings of the The 17th Annual Workshop of the Australasian Language Technology Association, pp. 120–125 (2019)

[16] Zhao, Y., et al.: Design and Implementation of the Move Recognition System for Fund Project Abstract. Information studies: Theroy & Application 45(8), 162–168 (2022)
[17] Yu, G. H., et al.: Masked Sentence Model Based on BERT for Move Recognition in Medical Scientific Abstracts. Journal of Data and Information Science 4(4), 42–55 (2019)
[18] Cohan, A., et al.: Pretrained Language Models for Sequential Sentence Classification. arXiv preprint arXiv 190904054v2 (2019)
[19] Sollaci, L.B., et al.: The introduction, methods, results, and discussion (IMRAD) structure: a fifty-year survey. J Med Libr Assoc 92(3), 364–367 (2004)
[20] Haynes, R.B., et al.: More informative abstracts revisited. Ann Intern Med 113(1), 69–76 (1990)
[21] Hayward, R.S., et al.: More informative abstracts of articles describing clinical practice guidelines. Ann Intern Med 118(9), 731–737 (1993)
[22] Nakayama, T., et al.: Adoption of structured abstracts by general medical journals and format for a structured abstract. J Med Libr Assoc 93(2), 237–242 (2005)
[23] Kulkarni, H.: Structured abstracts: still more 124(7), 695–696 (1996)
[24] Hopewell, S., et al.: CONSORT for reporting randomized controlled trials in journal and conference abstracts: explanation and elaboration. PLoS Med 5(1), e20 (2008)
[25] Manning, C.D., Surdeanu, M., Bauer, J., et al.: The Stanford CoreNLP natural language processing toolkit. In: Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations, pp. 55–60 (2014)
[26] Andrade, C.: How to write a good abstract for a scientific paper or conference presentation. Indian J Psychiatry 53(2), 172–175 (2011)
[27] Abdollahpour, Z., et al.: Rhetorical Structure of the Abstracts of Medical Sciences Research Articles. La Prensa Medica Argentina 105(2), 1–5 (2019)
[28] Hirohata, K., et al.: Identifying Sections in Scientific Abstracts using Conditional Random Fields. In: Proceedings of the Third International Joint Conference on Natural Language Processing, pp. 381–388 (2008)
[29] Neumann, M., King, D., Beltagy, I., et al.: ScispaCy: fast and robust models for biomedical natural language processing. arXiv preprint arXiv 1902.07669 (2019)
[30] Carletta, J.: Assessing Agreement on Classification Tasks: The Kappa Statistic. Computational Linguistics 22(2), 249–254 (1996)
[31] Zhang, Z.X., et al.: Identifying Moves of Research Abstracts with Deep Learning Methods. Data Analysis and Knowledge Discovery 3, 1–9 (2019)
[32] Zhang, Z.X., et al.: Moves Recognition in Abstract of Research Paper Based on Deep Learning. In: 2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL), pp. 390–391 (2019)
[33] Vaswani, A., et al.: Attention is all you need. arXiv preprint arXiv 170603762 (2017)
[34] Zhang, Z.X., et al.: Building an Artificial Intelligence Engine Based on Scientific and Technological Literature Knowledge. Journal of Library and Information Science in Agriculture 33(1), 17–31 (2021)

AUTHOR BIOGRAPHY

Jie Li is a PhD student at the National Science Library, Chinese Academy of Sciences, Beijing, China; Department of Library, Information and Archives Management, School of Economics and Management, University of Chinese Academy of Sciences, Beijing, China; Her research interests include deep clustering, unsupervised learning, reviewer recommendation, journal recommendation and corpus construction.
Email address: lijie201909@mail.las.ac.cn
ORCID: 0000-0002-2252-1865

Gaihong Yu received her Ph.D. degree from University of Chinese Academy of Sciences, Beijing, China, in 2019. She is currently an associate research librarian of National Science Library, Chinese Academy of Sciences, Beijing, China. Her main fields of interest are scientific and technological intelligence monitoring, intelligent information processing, and deep semantic mining of scientific and technological literature content.
Email address: yugh@mail.las.ac.cn
ORCID: 0000-0003-1301-2871

Zhixiong Zhang is Deputy Director of the National Science Library, Chinese Academy of Sciences, Research Librarian (Level II), Doctor, Doctoral Supervisor. His main research fields are deep learning technology methods, semantic annotation, information extraction, network science and technology information monitoring, and preprint academic exchange. Winners of the National Hundred and Ten Thousand Talents Project, the "Distinguished Researcher Program of the Chinese Academy of Sciences", and the "Pollyanna Chu Excellent Teacher Award of the Chinese Academy of Sciences". He is currently the Director of the Information Technology Professional Committee of the Chinese Society for Science and Technology Information, the Deputy Chairman of the National Library Standardization Technical Committee, and the Deputy Director of the Knowledge Organization Professional Committee of the Chinese Society for Science and Technology Information; Co editor in chief of Data Intelligence (DI) journal, deputy editor in chief of Data Analysis and Knowledge Discovery journal, editorial board member of Journal of Data and Information Science (JDIS), Digital Library Forum, Think Tank Theory and Practice, and Intelligence Engineering journal. Published one monograph, over 150 research papers, and three translated works. Hosted and participated in over 40 national and provincial level projects.
Email address: zhangzhx@mail.las.ac.cn
ORCID: 0000-0003-1596-7487