Master's Thesis in Graduate School of
Library, Information and Media Studies

Automatic Generation of Research Objectives and Citation Sentences in Academic Papers

March 2021

201821629
MAENO SHOHEI

Automatic Generation of Research Objectives and Citation Sentences in Academic Papers
学術論文の研究目的と引用文の自動生成

Student No.: 201821629
氏名： 前野　翔平
Name: MAENO SHOHEI

The purpose of this study is to automatically generate the objective sentences and citation sentences included in the introduction secetion of academic papers. The introduction of an academic paper is the most important section in order to explain the importance of research. Therefore, it is a challenge to write an introduction that will help the reader understand the significance of the research. In addition, the introduction is a section that always contains sentences that state research objectives and citations of related researches, regardless of the research field. Therefore, we believe that the automatic generation of objective sentences and citation sentences can make the writing of the introduction section easier.

To build the academic article dataset used in this study, we manually collected open access articles. For the automatic generation of objective sentences, we first built the objective sentence extraction model to collect automatically the objective sentences from articles. Second, we used the Method, Results, Discussion and similar sections of articles, and trained the model to generate the objective sentence. For the automatic generation of citation sentences, we first collected citation sentences from the introduction section of colleted papers and extracted abstracts of the corresponding citation sentences from PubMed (`https://pubmed.ncbi.nlm.nih.gov/`) to build the dataset. Secondly, we trained the model to generate citation sentences from the abstract of the cited paper and the Method, Results, Discussion, and similar sections of the collected paper.

We implemented the ROUGE-TITLE and ROUGE-ABSTRACT metrics to determine whether the quality of human-written or model-generated sentences is better. ROUGE-TITLE calculates the ROUGE-1 of the title of the paper and the sentence, and ROUGE-ABSTRACT calculates the ROUGE-1 of the abstract of the paper and the sentence. In the objective sentence generation task, the title and abstract of collected papers are used to calculate the two metrics. On the other hand, the title and abstract of the cited papers are used in the citation sentence generation task. ROUGE-TITLE, ROUGE-ABSTRACT of human-written sentences and model-generated sentences are computed to evaluate the models.

We succeeded in generating equally or higher quality sentences than humans in the objective sentence generation task and the citation sentence generation task with a probability of 17.5% and 11.8%, respectively. The automatic generation of the objective sentences and citation sentences in the introduction section will help people who have difficulties in writing papers in English.

Principal Academic Advisor: Tetsuya MAESHIRO
Secondary Academic Advisor: Hideo JOHO

Automatic Generation of Research Objectives and Citation Sentences in Academic Papers

MAENO SHOHEI

Graduate School of Library, Information and Media Studies

University of Tsukuba

March 2021

Contents

Chapter 1

Introduction

The aim of this study is to automatically generate the research objective sentence and the citation sentence written in the introduction section of an academic paper. By applying the deep learning techniques and the concept of sentence summarization, the two sentence generation tasks in this study, i.e., the objective sentence generation task and the citation sentence generation task, are realized. The implementation of the two sentence generation tasks in this study can be useful for generating drafts of objective sentences and citation sentences, and for helping researchers who are not accustomed to writing papers in English.

The introduction is the section of the paper that contains the most of the researcher's opinion about the significance of the study. By reading the introduction, the readers can understand why the research is conducted. Therefore, no matter how valuable the author thinks the research is, it may not adequately convey the original value of the research if the introduction is poorly written. In addition, as the research field becomes broader and the growing number of papers are published[29], there is a possibility that it will be reviewed by people who are not experts in the research field, resulting in the need of the introduction written in an easy-to-understand manner. Therefore, we should pay more attention to the importance of the introduction.

Due to the importance of the introduction as mentioned above, the automatic generation of introductions will help to reduce the burden of writing papers. However, since the length of an introduction section exceeds several hundred words, it is difficult to generate the full text of an introduction, even using language generation models that have shown high performance in various language generation tasks[16]. In order to cope with such problems, language generation models have been improved in recent years to be able to handle even longer sentences[19]. However, we will treat the sentences in the introduction as separate language generation tasks by decomposing them into objective sentences and citation sentences. Therefore, we will work on a new language generation task: automatic generation of the objective sentence and the citation sentence in the introduction.

Regardless of the field of research, the objective sentence and citation sentence are always included in the introduction. In some papers, these two sentences alone make up more than half of the introduction. Therefore, the automatic generation of these two sentences can greatly reduce the burden of writing an introduction. The automatic generation of the citation sentences is worth working on because it contributes to reducing the time required for literature research. Futher, there is no doubt that research goal is the most important concept in the process of conducting research and should always be made clear. However, that does not necessarily mean that the objective sentence in the paper is written in a clear and understandable manner. For example, in the following sentence *"Thus, we attempted to solve these problems."*[24], it is difficult to understand the purpose of the research without referring to the preceding and following sentences. On the other hand, the following sentence *"To understand the function of ARRDC3 we used overexpression and short hairpin RNA-mediated downregulation in human breast cancer cells."*[28] condenses the purpose of the research into one sentence, which is directly and clearly conveyed to the reader. Therefore, the automatic generation of the objective sentence, which consists of only one sentence, not only emphasizes the research purpose more, but also reduces the burden when people read the paper.

Although many language generation tasks have been studied[27], there is still a lack of research on sentence generation to make writing academic papers easier. Still, there have been studies on building huge datasets of academic papers[8] and on automatic generation of abstracts[2, 3]. However, we have not found any research that has addressed sentence generation to reduce the burden of writing the introduction section. Therefore, this research can be expected to be a catalyst for more active research on automatic generation of academic papers.

Chapter 2

Related Work

Research has been conducted to support the writing of academic papers using language generation techniques. For example, Demir et al. performed automatic generation of academic papers in LaTex format[1]. They collected 799 LaTex files of computer vision papers from the arXiv (`https://arxiv.com/`) and automatically generated LaTex sources on a character-by-character basis. They used a deep learning model, but the number of data is small. In addition, they evaluated the results generated by multiple models, and did not evaluate whether the generated papers were better than those written by humans or not. In our research, we evaluated the model by generating sentences using more than hundreds of thousands of papers and evaluating them against sentences written by humans.

There is also a lot of research on automatic generation specific to the elements and sections that make up a paper. For example, Wang et al. developed a model that automatically generates abstracts from titles and automatically modifies the generated abstracts[2]. Hu et al. calculated the relevance of the sentences extracted from the Abstract, Introduction, and Conclusion of the paper and the sentences in the Abstract of the references, and automatically generated the Related Work section with the sentences with high relevance[4]. Meng et al. used the Sequence2Sequece model to automatically generate keywords for research[5]. As described above, there have been studies on automatic generation of abstracts, related research sections, and keywords, but there have been no studies focusing on introductions.

Xing et al. generated the corresponding citations from the sentences before and after the citations in the related research section and the abstract of the cited reference[6]. However, we used Method, Results, and Discussion as the citation information. This makes it possible to generate citation sentences while writing the specific sections of the research such as Method, Results, Discussion.

As described above, research on the automatic generation of academic papers is active, but the task of language generation is still lacking compared to research on language generation such as translation, question answering generation, and dialogue systems. This is because the available datasets of academic papers are not well constructed. Many researches have been working on language generation by collecting dozens of articles from arXiv or specific websites. Article datasets consisting of tens of thousands of data are ACL Anthology[7] and S2ORC[8]. The two datasets consist of text obtained by analyzing PDFs of articles, but the PDF format varies from article to article, making it impossible to extract the correct article data. In addition, S2ORC includes non-refereed papers, which means that some papers are of low quality. The details are described in Section 5.1.

Chapter 3

Task Definition

In this chapter, we provide definitions for the two sentence generation tasks we will perform in this study.

3.1 Objective Sentence Generation Task

The first task of this study is the automatic generation of the objective sentence, that is, the sentence that corresponds to the research objective written in the introduction. The research objective statement is one of the most important statements in the paper, and it condenses the research content. Therefore, the objective sentence generation task can be treated as a one-sentence summary task of the research content.

Summarization is the extraction of important sentences from a concrete text or the generation of new sentences that incorporate important elements of the original text[26]. Applying the concept of summarization, it is possible to generate a objective sentence from the detailed contents of a paper.

An academic paper basically consists of sections such as Abstract, Introduction, Method, Results, Discussion, and Conclusion. These sections can then be categorized into sections with concrete content and sections with abstract content. The former section contains Method, Results, and Discussion, and the latter section contains Abstract, Introduction, and Conclusion. By applying the mechanism of text summarization, it is possible to generate a section with abstract contents from a section with concrete contents of an article. Therefore, the purpose of this study is to automatically generate the objective sentences in the Introduction from the sections of the paper that contain detailed contents such as Method, Results, and Discussion and similar sections.

By including Abstract and Conclusion in the input data, it can be assumed that the objective sentence can be generated more easily. This is because those sections contain the objective sentence, which makes the language generation model easier to learn. However, even if a high-quality objective sentence is generated by training a model with an Abstract and Conclusion in the input data, the Abstract and Conclusion must be prepared when the model is actually used by humans in real situations. In that case, humans would have an extra burden because they would have to write sections with abstract content. Therefore, when it is possible to generate objective sentences from sections with specific content, the objective sentences can be generated dynamically from only those sections that are less burdensome to write.

3.2 Citation Sentence Generation Task

The second task of this study is the automatic generation of the citation sentences in the introduction. A citation sentence is a statement that condenses the contents of a cited reference into a single sentence. Therefore, it is possible to automatically generate citation sentences from the contents of the reference.

We use abstracts as the content of references in our citation generation task. The abstract is a short summary of the entire paper, with one or two sentences each describing the purpose of the research, methods, results, and contributions. Therefore, by using Abstract, it is possible to handle the important sentences of the entire paper in a pseudo-structure. It is also required to generate citations that reflect the research content of the papers citing the refenrences, rather than generating sentences that only reflect the content of the cited paper. In the Introduction, the citation sentence relates the content of the paper citing the reference to the content of the cited reference. Therefore, the goal of the citation generation task is to generate a citation sentence from the Method, Results, and Discussion of the paper citing the reference and the Abstract of the cited reference.

Chapter 4

Method

4.1 Data Acquisition

We will independently collect the academic papers that will be dealt with in this study. Several datasets containing academic papers have been proposed. A number of studies on text generation for academic papers have used the ACL Anthology Network as a dataset for academic papers. The ACL Anthology Networkc[7] is built up of papers from international conferences and journals sponsored by the ACL. The ACL has many high quality papers because there are many international conferences whose acceptance rate is less than 30%. The number of papers is 23,766 as of December 23, 2020. However, it is possible that the quality of the papers is high but the quality of the data is low. Because to build the ACL Anthology Network, PDF parsing was performed to extract the text of the articles. PDF parsing may extract the contents of the same section as different sections if the pages are different. As a result, it is difficult to extract the paper data accurately without loss.

Another large dataset of academic papers called S2ORC has been constructed[8]. This dataset collects 8.1M open access articles, which is the largest dataset among academic article datasets. However, the quality of the papers varies because some papers are not peer-reviewed. In addition, like the ACL Anthology Network, the content of the articles may be flawed due to PDF parsing. Therefore, in this study, we constructed our own dataset of academic papers that overcomes the problems of existing datasets.

We only deal with Open Access articles from journals published by Nature (`https://www.nature.com/`). Instead of extracting the article data by PDF parsing, the text data is extracted by analyzing the HTML structure of the article that can be obtained from the journal website. This allowed us to successfully extract the full text of the title, abstract, introduction, methods, and reference list without loss. In the end, we obtained 160,021 open access articles from 88 journals as shown in Table 4.1. Although the number of data is smaller than that of S2ORC, we obtained more data than the ACL Anthology Network. Furthermore, the quality of the papers should be high because many journals have a high impact factor as listed in Table 4.2. In addition, the accurate extraction of the paper data made the quality of the data high. Therefore, we have successfully solved the problems of the existing academic paper dataset and constructed a new academic paper dataset.

Table 4.1: Article dataset

Item	Quantity
Articles	160,021
Journals	88

Table 4.2: Journal data

Name	Articles	H-index
Nature Genetics	1	550
Oncogene	875	329
Nature Communications	26,670	298
British Journal of Cancer	230	224
International Journal of Obesity	265	218
Neuropsychopharmacology	243	214
Cell Death and Differentiation	203	208
Molecular Psychiatry	525	207
Leukemia	408	185
Scientific Reports	113,737	179
Gene Therapy	93	156
Cell Research	256	153
European Journal of Clinical Nutrition	148	152
Laboratory Investigation	36	147
Modern Pathology	35	145
Bone Marrow Transplantation	62	125
European Journal of Human Genetics	308	122
Genetics in Medicine	2	121
Heredity	155	115
Spinal Cord	63	104
Genes and Immunity	48	96
Cell Death and Disease	1,584	96
Eye	61	93
Journal of Human Hypertension	65	93
Mucosal Immunology	105	92
Journal of Perinatology	54	89
Journal of Exposure Science and Environmental Epidemiology	49	87
Cancer Gene Therapy	68	84
Hypertension Research	55	84
International Journal of Impotence Research	11	81
Light: Science and Applications	623	80
Journal of Human Genetics	56	79
British Dental Journal	12	77
Experimental and Molecular Medicine	569	75

Continued on next page

Table 4.2 – *Continued from previous page*

Name	Articles	H-index
Translational Psychiatry	1,961	74
Cellular and Molecular Immunology	35	74
Polymer Journal	15	66
NPG Asia Materials	559	65
Prostate Cancer and Prostatic Diseases	60	61
Scientific Data	79	48
Blood Cancer Journal	472	44
Nature Ecology and Evolution	1	43
Oncogenesis	584	39
International Journal of Oral Science	50	39
Bone Research	145	34
Nutrition and Diabetes	377	34
Horticulture Research	438	27
npj Computational Materials	368	25
npj Quantum Information	315	24
npj Quantum Materials	249	23
Microsystems and Nanoengineering	251	23
Cell Discovery	230	22
npj Biofilms and Microbiomes	142	21
Cell Death Discovery	376	21
npj Primary Care Respiratory Medicine	247	21
npj Schizophrenia	116	19
Signal Transduction and Targeted Therapy	134	18
npj 2D Materials and Applications	141	18
npj Genomic Medicine	136	17
npj Breast Cancer	165	16
npj Aging and Mechanisms of Disease	55	16
npj Microgravity	109	15
npj Vaccines	187	15
npj Systems Biology and Applications	141	14
Palgrave Communications	462	13
Communications Biology	1,100	13
Communications Physics	388	11
Communications Chemistry	321	10
Human Genome Variation	77	10
npj Regenerative Medicine	54	8
BDJ Open	79	2
npj Science of Food	43	2
Spinal Cord Series and Cases	24	1
The Journal of Antibiotics	24	0
Communications Earth and Environment	16	0

Continued on next page

Table 4.2 – *Continued from previous page*

Name	Articles	H-index
npj Climate and Atmospheric Science	111	0
npj Clean Water	46	0
npj Science of Learning	49	0
Communications Materials	55	0
Humanities and Social Sciences Communications	68	0
npj Flexible Electronics	73	0
npj Precision Oncology	77	0
The Pharmacogenomics Journal	79	0
npj Materials Degradation	102	0
npj Parkinson's Disease	108	0
npj Digital Medicine	195	0
The ISME Journal	657	0
NO TITLE	1	0

4.2 Preprocessing

The article data collected from the Nature website contains unnecessary information and data. In this section, we describe the process to clean the data by removing those extra information and data.

4.2.1 Abbreviation Replacing

Words and phrases that occur frequently in academic papers are treated as abbreviations after they first appear in the paper. For example, in the following sentence, the word sequence *"least microenvironmental uncertainty principle"* will be abbreviated as *"LEUP"* from the next occurrence.

> *Here we propose the least microenvironmental uncertainty principle (LEUP) that may serve as a generative model of collective migration without precise incorporation of full mechanistic details.*[20]

Since the definitions of abbreviations vary from author to author, different word sequences may be treated as the same abbreviation in different papers. The following two sentences provide concrete examples. In the sentence above, *"SEM"* stands for *"scanning electron microscopy"*, but in the sentence below, *"SEM"* stands for *"standard error of the mean"*. Abbreviations play a role only in the defined paper, so when dealing with multiple papers, it is necessary to convert the abbreviations into the original word sequence.

> *The morphology was investigated using scanning electron microscopy (SEM) and atomic force microscopy (AFM).*[22]

> *All values are presented as mean ± standard error of the mean (SEM).*[23]

The AbbreviationDetector[9] is used to extract the correct names of abbreviations from the articles. We combine the title, abstract, and body sections of a paper into a single text and parse it using AbbreviationDetector to obtain the abbreviations and definitions in that paper in key-value format. Finally, we convert the matched abbreviations to the original word sequence by means of a regular expression.

4.2.2 Section Filtering

In this study, only the Method, Results, and Discussion, and similar sections not the full text of the articles collected from the Nature website, are treated in the objective sentence generation task and citation sentence generation task. The sections of the paper can be classified into two categories: those consisting of abstract content and those consisting of detailed content. The Introduction is included in the section with abstract content, and the Method, Results, Discussion, and similar sections are included in the section with concrete content. Sentence summarization generates abstract sentences from sentences that contain concrete information. Therefore, in this study as well, the objective sentences of Introduction are generated from specific sections. Abstract and Conclusion are the sections that contain the objective sentence. We exclude Abstract and Conclusion because the model can refer to the answer if there is an objective statement in the model input. The citation should be generated from the Abstract of the cited paper in addition to the Method, Results, and Discussion of the paper citing the reference. The Nature article contains sections that are not directly related to the content of the article or that do not contain important content, so unnecessary sections should be excluded. Table 4.3 shows a list of sections to be excluded.

Table 4.3: Removed sections

Title
references, supplementary material, conclusion, acknowledgment, supporting online material, outlook, concluding, summary, note added in proof, note in proof, data availability, abbreviation, notes, usage note, data archiving, author contribution, rights and permissions, related work, availability of data and material, comments, data and material availability, appendix, code and data availability, availability of data, publisher, abstract, introduction

4.2.3 Text Preprocessing

Text preprocessing plays a very important role in natural language processing. And the preprocessing changes depending on the task. For example, in a text classification task, words with high frequency of occurrence in all texts are removed. This study, however, addresses the language generation task. Prepositions and conjunctions, which occur more frequently, are also important elements in the construction of natural sentences. Therefore, the text preprocessing done in this study does not involve preprocessing to remove specific parts of speech. In natural language processing, depending on the task, the text is often divided into sentence units. However, some software that splits text into sentences has problems such as splitting by phrases instead of sentences, or splitting by periods other than the period at the end of a sentence. The following is a list of preprocessing steps performed in this study.

- Remove URL
- Remove words longer than 45 characters
- Remove Greek letters and symbols used in mathematical formulas
- Convert uppercase to lowercase
- Remove single letters
- Remove parentheses and their contents
- Remove periods in fig. figs. ref. refs. et al. e.x. i.e.
- Set all number words to 0
- Split the main text into sentences using NLTK's Sentence Tokenizer[10]
- Exclude sentences with less than 10 words

4.3 Basic Datasets

In this section, we describe the construction of the basic dataset that will be used in the objective sentence generation and citation sentence generation tasks.

4.3.1 Objective Sentence Generation Dataset

Since the objective sentence generation task is a new task in the sentence generation task, we need to build our training dataset. The dataset consists of pairs of sections with the detailed content of the paper and the objective sentence. The sections with detailed content are Method, Results, Discussion, and similar sections. The objective sentence is the sentence that corresponds to the research aim contained in the introduction. Although it is possible to manually extract the objective sentences by reading the introduction, performing it from 160,021 articles would require an enormous amount of time. Therefore, in order to extract the objective sentence more efficiently, we build an objective sentence extraction model.

The first step is to prepare the dataset to be used in training the model. Among the collected Nature articles, we manually extracted the objective sentences from the Introduction for 1000 articles from journals with high SJR. The calculation of SJR is based on Scimago Journal & Country Rank (`https://www.scimagojr.com/`). We thought that the higher the level of the journal, the clearer the objective sentence and the higher the quality.

As a result of manual extraction, 1038 objective sentences were extracted from 1000 papers. We also randomly extracted 1038 sentences from the Introduction that were neither the objective sentence nor the citation sentence. In the end, 2076 sentences were extracted to build the objective sentence extraction model as shown in Table 4.4.

Table 4.4: Dataset for the objective sentence extraction model

Item	Quantity
Objective sentences	1,038
Non-objective sentences	1,038

To build the model, we used the pre-trained BERT model for fine tuning[11]. By using this model, it is possible to determine whether a sentence fed into the model is the objective sentence or not.

During training, we set the maximum length of sentences to be input to the model to 30, the dropout to 0.4, and we used AdamW as the optimization function[12] and the learning rate to 2e-5, and conducted four epochs of training. We split 80%, 10%, and 10% of the dataset into training, validation, and test data, respectively. As explained in Table 4.5, the test result shows that Accuracy, Precision, Recall, and F1 are 0.938, 0.910, 0.971, and 0.940, respectively.

We used the objective sentence extraction model constructed here to automatically extract objective sentences from the Introduction of a paper. The output from the model indicates the probability that the input sentence will be identified as the objective sentence. We set a threshold to 0.8 for the probability that a sentence fed into the model is identified as the objective sentence. However, some of the sentences identified as objective sentences are in fact not suitable as objective sentences. For example, the sentence *"Therefore, we undertook* *a multicenter retrospective study to understand these aspects better and develop a more targeted plan for the future."*[24] contains the pronoun "these" and it is impossible to convey the purpose of the research in this sentence alone. In our objective sentence generation task, the goal is to generate an objective sentence that can be used as a research objective with only one sentence. For this reason, we exclude sentences that contain pronouns such as "these", "their", "them", and "those", even if they are identified as objective sentences. In addition, some of the sentences identified as objective sentences have few technical terms and are too abstract. For example, the sentence *"In the study presented here, we tested these hypothesis"*[25] does not include the keywords of the research and lacks specificity. For this reason, we exclude sentences that contain more than 75% of the words that Japanese junior high school students learn in middle school. We reference those vocabularies from Weblio (`https://ejje.weblio.jp/`) and Table 4.6 shows a list of words to be excluded. Also, the objective sentence is as important as the title in the paper. Therefore, we exclude sentences that do not contain any of the words in the title.

Table 4.5: Results of testing the objective sentence extraction model

Accuracy	Precision	Recall	F1
0.938	0.910	0.971	0.940

Table 4.6: List of vocabularies that junior high school students in Japan learn

Vocabulary
a, ability, able, about, above, accept, account, across, act, action, activity
actually, add, address, afraid, after, afternoon, again, against, age, ago, agree
air, airplane, all, allow, almost, alone, along, already, also, although, always
am, among, amount, and, angry, animal, another, answer, any, anything, appear
apple, area, arm, around, arrive, art, article, artist, as, ask, at, attack, attention
aunt, autumn, away, baby, back, bad, bag, ball, base, baseball, basket, basketball
bath, be, beautiful, because, become, bed, before, begin, behind, believe, bell
below, beside, between, beyond, bicycle, big, bird, birthday, black, block, blood
blue, board, boat, body, book, both, bottle, bottom, box, boy, break, break, down
breakfast, bridge, bright, bring, brother, brown, build, bus, business, busy
but, butter, buy, by, cake, call, camera, can, cap, car, card, care, carry, case
cat, catch, cause, center, century, certain, chair, challenge, chance, change
cheap, check, child, choice, choose, church, circle, city, class, classroom, clean
clear, climb, clock, close, cloud, club, coat, coffee, cold, color, come, common
company, complete, condition, consider, contact, continue, control, cook, cool
copy, corner, correct, cost, could, count, country, couple, course, court, cover
cow, cream, create, cross, crowd, cry, culture, cup, cut, daily, dance, dangerous
dark, date, daughter, day, dead, deal, dear, death, decide, decision, deep, department
describe, desk, develop, diary, dictionary, die, difference, different, difficult
dinner, dirty, discover, disease, dish, distance, do, doctor, dog, doll, door

Continued on next page

Table 4.6 – *Continued from previous page*

Vocabulary
down, dream, dress, drink, drive, drop, dry, during, each, ear, early, earth
east, easy, eat, edge, effect, effort, egg, eight, eighteen, eighty, either, eleven
else, empty, end, enemy, enjoy, enough, enter, especially, even, evening, event
ever, every, everything, example, except, exercise, expect, experience, explain
eye, face, fact, fall, family, famous, far, farm, farmer, fast, fat, father, fear
feel, few, field, fifteen, fifth, fifty, fight, figure, fill, fill, in, film
find, fine, finger, finish, fire, first, fish, five, fix, flat, floor, flower
fly, follow, food, foot, football, for, force, foreign, forest, forget, form
former, forty, forward, four, fourteen, fourth, free, fresh, friend, from, front
fruit, full, fun, future, game, garden, gas, gate, general, gentleman, get, girl
give, glad, go, go, on, god, gold, good, good, night, government, grade, grass
great, green, group, grow, guess, guide, gun, hair, half, hand, happen, happy
hard, hat, have, he, head, health, healthy, hear, heart, heat, heavy, help, her
here, hers, hi, high, hill, him, himself, his, history, hit, hold, hole, holiday
home, hope, horse, hospital, hot, hotel, hour, house, how, however, human, hundred
hungry, hurt, husband, ice, idea, if, ill, image, important, improve, in, include
increase, individual, industry, information, inside, instead, interest, international
into, invite, iron, island, it, its, itself, job, join, jump, junior, just, keep
key, kid, kill, kind, king, kitchen, knife, know, knowledge, lady, lake, land
language, large, last, late, later, laugh, law, lead, leader, learn, least, leave
leg, length, less, lesson, let, letter, level, library, lie, life, light, like
likely, line, list, listen, little, live, local, long, look, look, up, lose, lot
love, low, lunch, machine, magazine, mail, main, major, make, man, many, map
mark, market, master, material, matter, may, maybe, me, meal, mean, measure, meat
medical, medicine, meet, meeting, member, memory, message, method, middle, might
milk, million, mind, mine, minute, miss, mistake, modern, moment, money, month
moon, morning, mother, mountain, mouth, move, movie, mr, much, music, must, my
myself, name, narrow, nation, national, natural, nature, near, nearly, necessary
need, neither, never, new, news, newspaper, nice, night, nine, no, noise, none
noon, nor, north, nose, not, note, notebook, nothing, notice, now, object, ocean
of, off, offer, office, official, often, oh, oil, old, on, once, one, only, open
opinion, or, order, other, our, ours, out, outside, over, own, page, pain, paint
pair, paper, parent, park, part, particular, party, pass, past, patient, pay
peace, pen, pencil, people, perhaps, period, person, personal, phone, piano, pick
pick, up, picture, piece, place, plan, plane, plant, play, player, please, pocket
point, police, political, poor, popular, population, position, possible, post
pot, power, practice, prepare, present, president, pretty, price, private, probably
problem, process, produce, product, program, promise, protect, provide, public
pull, purpose, put, put, on, quarter, queen, question, quick, quickly, quiet
quite, race, radio, rain, raise, rate, rather, reach, read, ready, real, realize
really, reason, receive, recent, recently, record, red, remain, remember, report

Continued on next page

Table 4.6 – *Continued from previous page*

Vocabulary
require, rest, restaurant, result, return, ride, right, ring, rise, river, road
rock, room, round, rule, run, sad, safe, salt, same, save, say, scene, school
science, sea, season, seat, second, see, seem, sell, send, sense, serious, serve
service, set, up, seven, seventy, several, shall, shape, share, she, shine, ship
shirt, shoe, shop, short, should, shout, show, shut, sick, side, sight, sign
silver, similar, simple, simply, since, sing, single, sister, sit, situation
six, sixteen, sixth, sixty, size, skin, sky, sleep, slow, slowly, small, smile
snow, so, social, society, soft, some, somebody, someone, something, sometimes
son, song, soon, sorry, sort, sound, south, space, speak, special, speech, speed
spend, sport, spread, spring, square, stand, standard, star, start, state, station
stay, step, still, stone, stop, store, story, straight, strange, street, strong
student, study, style, subject, success, such, suddenly, sugar, suggest, suit
summer, sun, support, suppose, sure, surprise, sweet, swim, system, table, take
take, off, talk, tall, tax, tea, teach, teacher, team, tear, telephone, television
tell, ten, tennis, term, test, than, thank, that, the, their, them, themselves
then, there, these, they, thick, thin, thing, think, third, thirteen, thirty
this, those, though, thousand, three, through, throw, ticket, till, time, to
today, together, tomorrow, tonight, too, tooth, top, touch, toward, town, trade
train, travel, tree, trip, trouble, true, try, turn, twelve, twenty, two, type
uncle, under, understand, university, until, up, us, use, useful, usually, valley
value, various, very, view, village, visit, voice, wait, walk, wall, want, war
warm, wash, watch, water, wave, way, we, weak, wear, weather, week, weekend, welcome
well, west, western, wet, what, when, where, whether, which, while, white, who
whole, whom, whose, why, wide, wife, wild, will, win, wind, window, winter, wish
with, within, without, woman, wonderful, wood, word, work, world, worry, would
write, writer, wrong, yardz, year, yellow, yes, yesterday, yet, you, young, your

The data to be handled in the objective sentence generation task should have a total of at least 1,000 words in Method, Results, Discussion and similar sections, and at least 15 words and no more than 50 words in the objective sentence. In the end, we acquired 158,035 data as our objective sentence generation dataset. As explained in Table 4.7, the average total word count for Method, Results, Discussion and similar sections is 3,704 words, and the average word count for the objective sentence is 29 words.

Table 4.7: Statistics of objective sentence generation dataset

Item	Value
Data	158,035
Mean length of Main text	3,704
Mean length of Objective sentence	29

4.3.2 Citation Sentence Generation Dataset

To build the dataset for the citation sentence generation task, we first extracted the citation sentences from Introduction. When we extracted the article data from the Nature journal website as described on Data Acquisition, the citation sentences are tagged with *REFERENCE* along with the number of the cited reference as shown below.

> *This results in a variety of models specific to certain individual species<REFERENCE/1,3,4/>.*[20]

Therefore, it is easy to extract citation sentences from the Introduction using regular expressions. Note that there are three main types of citation sentences.

1. Sentences that cite more than one point from just one reference

   > *Collective migration has been achieved, for example, through the introduction of a ferromagnetic-like interaction potential, which locally aligns particle velocities polarly, or a liquid-crystal-like interaction potential, which aligns particle velocities nematically<REFERENCE/7/>.*[20]

2. Sentences with multiple points from multiple references

   > *Collective movement of dense populations is observed in several biological systems at different scales, from massive migration of mammals<REFERENCE/1/> to cells during embryogenesis<REFERENCE/2/>.*[20]

3. Sentences with a single point from multiple references

   > *This results in a variety of models specific to certain individual species<REFERENCE/1,3,4/>.*[20]

In this study, only the citation sentences 1 and 3 above are extracted from the Introduction. The model to be trained for the citation sentence generation task has two inputs: the Method, Results, Discussion and similar sections of the paper citing the reference, and the Abstract of the cited reference. However, when dealing with the citation sentences 2, the number of inputs to the model will increase with the number of points cited. This makes it impossible to unify the design of the model. On the other hand, in case 1, there is only one point cited, regardless of where it is cited in a sentence, so there is only one pair of papers citing the reference and cited reference. In the case of 3, there is only one point cited, so there can be as many pairs of papers citing references and cited references. In the case of the example sentence 3, three pairs of two inputs are created: the paper citing the reference and the reference cited.

In this study, the abstracts of the cited references are extracted from PubMed (`https://pubmed.ncbi.nlm.nih.gov/`) using BioPython[13]. Therefore, only citations cited from articles published in PubMed are included in the citation sentence generation dataset.

The data to be handled in the citation sentence generation task should be a total of at least 1,000 words in the Method, Results, Discussion and similar sections of the article citing the reference, at least 100 words in the Abstract of the reference, and at least 15 words and no more than 50 words in the citation. In the end, we obtained 196,426 data as our citation sentence generation dataset. As explained in Table 4.8, the average total word count for Method, Results, Discussion and similar sections of the papers citing the reference is 4,072 words, while the average word count for the Abstract of the cited reference is 193 words and the average word count for the citation sentence is 29 words.

Table 4.8: Statistics of citatoin sentence generation dataset

Item	Value
Data	196,426
Mean length of Main text	4,072
Mean length of Cited Abstract	193
Mean length of Objective sentence	27

4.4 Final Dataset

We build Final Datasets from Basic Datasets in order to train models more efficiently. In Section 4.3, we constructed the objective sentence generation dataset and the citation sentence generation dataset. For each dataset, the average total number of words in the Method, Results, Discussion and similar sections of the articles collected from Nature is 3,704 and 4,072 words, respectively. A large number of words in a sentence input to the model increases the training time of the model and causes unstable training.

Therefore, we conjecture that a small number of distinctive sentences can be used to train a model efficiently. In order to solve this problem, we extract sentences from a text that concatenates Method, Results, Discussion and similar sections based on certain features. We then speculate that the extracted sentences can be used to train the model more efficiently. The following is a description of the different datasets consisting of sentences extracted based on the four features. Four datasets are constructed for each of the objective sentence generation and citation sentence generation datasets. For the citation generation dataset, we also extract sentences from the Abstract of the cited article, apart from the Method, Results, Discussoin and similar sections of the article citing the reference.

4.4.1 High Similarity Dataset

The sentences included in the first final dataset are those that have a high semantic similarity to the objective sentence or the citation sentence in each task. Therefore, we call this final dataset High Similarity Dataset.

For example, in the dataset for the objective sentence generation task, sentences with high semantic similarity to the objective sentence are extracted from the concatenated text of Method, Results, Discussion and similar sections. In the dataset for the citation sentence generation task, we extract sentences that have a high semantic similarity to the citation sentence from the Method, Results, and Discussion of the citing paper and the Abstract of the citing paper, respectively.

To calculate the similarity between two sentences, we use cosine similarity. The sentence vectors used in the calculation of cosine similarity are obtained using SentenceTransformer[15].

We hypothesize that if we train the model with sentences that have a high cosine similarity to the objective sentence or citatoin sentence on each generation tasks, the model will learn more efficiently to generate better sentences. This is because in training the neural network, the model is not trained by focusing only on the letters of the words, but as a vector of hundreds of dimensions. However, the extraction of sentences with high cosine similarity is only feasible when training the model. This is because, in fact, when a human wants to perform automatic objective sentence generation, the objective sentence does not exist, so the cosine similarity cannot be calculated in real situation. Therefore, we need to consider the characteristics of sentences that do not depend on the objective sentence or the citation sentence. It is the extraction of sentences using the average probability of occurrence of words.

4.4.2 High Average Probability of Word Occurrence Dataset

The sentences included in the second final dataset are the sentences with the higher average probability of word occurrence. Therefore, we define this final dataset as High Average Probability of Word Occurrence Dataset.

Important words and phrases related to the research topic will appear frequently in the paper. The objective sentence is one of the most important sentences in the paper, so it is likely to contain the keywords of the research. Therefore, we believe that sentences with high average probability of word occurrence are effective for training the objective sentence generation model.

Citation sentences are one-sentence summaries of the references. Therefore, it is likely to contain important keywords from the references. The citations in the introduction are also likely to play an important role in conveying the value of the research in the paper citing the reference. Thus, it is also probable that it will contain words that have a high frequency of occurrence in the papers citing the reference. Therefore, we believe that automatic generation of citation sentences is feasible by using sentences with a high average probability of word occurrence for training the model.

Only nouns, adjectives, and adverbs are considered in the calculation of word occurrence probability. Articles, prepositions, etc. are excluded because they do not make sense on their own, although they occur frequently. Verbs are also included in the sentence, but are not often included in the keywords of the research, so they are excluded. The probability of word occurrence is calculated by dividing the frequency of occurrence of nouns, adjectives, and adverbs in the text by the total number of noun, adjective, and adverb words in the text. Then, the average probability of a word occurring in a sentence is calculated by dividing the sum of the probabilities of noun, adjective, and adverb occurrences in the sentence by the total number of noun, adjective, and adverb words in the sentence.

4.4.3 High Score Dataset

The sentences included in the third final dataset are those with the higher harmonic average of the cosine similarity calculated in 4.4.1 and the average of the probability of occurrence of the words in the sentence calculated in 4.4.2. Therefore, we define this final dataset as High Score Dataset.

This dataset contains sentences that have a high similarity both semantically and superficially. Using harmonic mean prevents the extraction of sentences that are high in either the cosine similarity or the average probability of word occurrence. If sentences with high cosine similarity are extracted by using arithmetic mean, they will be indistinguishable from a High Similarity Dataset composed of sentences with high cosine similarity. For this reason, the harmonic mean, rather than the arithmetic mean, is more appropriate.

4.4.4 Random Dataset

The fourth final data set contains randomly selected sentences. Therefore, we define this final dataset as Random Dataset. In the objective sentence generation task, the sentences are randomly extracted from the text that concatenates Method, Results, Discusson and similar sections of the paper. In the citation sentence generation task, the sentences are randomly extracted from the Method, Results, and Discussion of the paper citing the reference and the Abstract of the cited reference, respectively. Since the previous three datasets extract sentences based on certain features, we infer that the model trained on this final dataset will be the worst performing model.

4.4.5 Definition of the Number of Extracted Sentences

Here, we explain the definition of the number of sentences to extract in the four final data sets that we have described so far.

First of all, in this study, when training a sentence generation model, we divide sentences into subwords and process them subword by subword, instead of dividing sentences into words and processing them word by word. If we divide sentences into words and process them on a word-by-word basis, the number of words that the model needs to handle increases as the number of articles increases. Then, the dense layer which is set as the output layer of the model will have hundreds of thousands of units, and the computational complexity will be huge. To solve this problem, a method of splitting a sentence into subwords has been proposed[14]. In this study, we also adopt the method of splitting a sentence into subwords. We used this method to reduce the vocabulary from about 270,000 words in the entire Nature article to 16,000 words.

For the final datasets other than the Random Dataset, the sentences with the highest values that were calculated in each of them are extracted until the total number of subwords in the extracted sentences exceeds 100. For Random Dataset, the sentences are extracted randomly until the total number of subwords in the extracted sentences exceeds 100 as well.

We set the threshold for the total number of subwords to 100 in order to reduce the amount of computation as much as possible. By setting a higher threshold, the model can learn a larger variety of vocabulary, which may result in a richer set of words in the generated sentences. Training a model using such a setup is left to the future experiments.

Chapter 5

Experiments

5.1 Generation Model

In the two sentence generation tasks we address in this study, we use Transformer[16] as our sentence generation model. Transformer is a model that has achieved state-of-the-art results on several existing sentence generation tasks. Therefore, we can expect high performance in both the objective sentence generation task and the citation sentence generation task. Transformer processes input tokens in parallel, rather than recursively as in the existing Sequence2Sequence model[30]. Therefore, the training time can be decreased than that of recursive neural networks.

In the objective sentence generation task, multiple sentences extracted from the concatenated text of Method, Results, Discussion and similar sections of the paper are input to the model, and Transformer is trained to output the objective sentence. On the other hand, in the citation sentence generation task, there are two inputs: the text of the paper citing the reference, and the text of the cited reference. For this reason, we have made one modification to the Transformer proposed in the original paper. We connect the Encoder that receives the text of the cited reference next to the Encoder that receives the text of the paper citing the reference. This allows the vector passed to the Decoder to include both the context of the paper citing the reference and the context of the cited reference. The following are the hyperparameters of Transfofmer configured for the two tasks. We set the number of dimensions of the model to 512, the number of dimensions of the feed forward network to 2,048, the number of heads of the self-attention to 8, and the number of layers of the Encoder and Decoder to 1.

5.2 Experimental Setup

In this study, we use byte-pair-encoding algorithm implemented from SentencePiece library to split a sentence into subwords[14]. The number of words in the vocabulary was set to 16,000. We used Adam as the optimization function[21] and set the initial learning rate to 1e-3. If the learning rate is constant, learning will stop in the middle of the process, so the learning rate is decayed by multiplying the learning rate by 0.95^{epoch} for each epoch. The final four datasets are divided into 80%, 10%, and 10% training data, validation data, and test data, respectively, and the model is trained for 100 epochs. The experiment was run on a GeForce RTX 3090.

When testing the trained model, the following procedure is applied to output the subwords and generate the sentences.

1. Output the subword with the highest probability of occurrence.
2. If the output subword is the same as a previously occurring subword, output the subword with the next highest probability.
3. When the first subword of a word overlaps with three previously output words, the subword with the next highest probability is output, excluding the previously
4. When the last three words output overlap, the subword with the next highest probability is output, excluding the subwords that make up the existing words.

Chapter 6

Results and Discussion

6.1 Evaluation Metrics

We used ROUGE-1, ROUGE-2, ROUGE-L[18], and BERTScore[17] to evaluate the trained models. For ROUGE-1, two types of ROUGE-1 are used: ROUGE-1(filtered), which limits the number of parts of speech, and ROUGE-1(unfiltered), which uses all parts of speech. ROUGE-1(filtered) calculates how close the generated sentence is to the correct sentence according to nouns, adjectives, and adverbs only. ROUGE-1(unfiltered) calculates how close the generated sentence is to the correct sentence, without distinguishing parts of speech. In most of the existing studies, ROUGE-1(unfiltered) has been used as a standard evaluation index. We believe that what is important in evaluating the objective sentence generation task is whether the keywords of the study are generated. Similarly, the key to evaluating the citation sentence generation task is whether the important keywords of the references are generated. Since prepositions and articles are meaningless by themselves, ROUGE-1(filtered) can be used to evaluate whether high-quality sentences are generated or not. ROUGE-2 calculates the similarity of correct and generated sentences in terms of bi-grams. ROUGE-L calculates the length of the longest common subsequence of correct and generated sentences. BERTScore calculates the average cosine similarity of words in the correct and generated sentences using the pre-trained BERT model. Since ROUGE only considers superficial agreement, we use BERTScore, which can consider the semantics of the words. The Final Score is the arithmetic mean of the values standardized across all test data for ROUGE-1(filtered), ROUGE-1(unfiltered), ROUGE-2, ROUGE-L, and BERTScore. The F1 value is used for ROUGE-1(filtered), ROUGE-1(unfiltered), ROUGE-2, ROUGE-L, and BERTScore.

All of the above evaluation metrics calculate how close the sentence generated by the model is to the sentence written by a human. Therefore, the evaluation is based on the assumption that human-written sentences are better than those generated by the model. Suppose that the sentences generated by the model are qualitatively better than those written by humans. To answer this question, we use ROUGE-TITLE and ROUGE-ABSTRACT for the objective sentence generation task and the citation sentence generation task, respectively. ROUGE-TITLE in the objective sentence generation task calculates ROUGE-1 for the title of the paper and the objective sentence written by humans, and for the title of the paper and the objective sentence generated by the model, for nouns, adjectives, and adverbs only, respectively. Since the title of the paper contains the keywords of the study, we can evaluate which sentence contains the more important keywords. ROUGE-ABSTRACT in the objective sentence generation task calculates ROUGE-1 for the abstract of the paper in which the objective sentence is written and human-written objective sentences, and for the abstract of the paper and model-generated objective sentences for nouns, adjectives, adverbs, respectively. For ROUGE-TITLE in the citation sentence generation task, the title to be treated is the title of the reference. ROUGE-ABSTRACT calculates ROUGE-1 for the abstract of the reference and human-written citations, and for the abstract of the reference and model-generated citations, for nouns, adjectives, and adverbs only, respectively.

In this study, we used as many high quality papers as possible as data, but we cannot objectively evaluate whether the objective sentences and citation sentences in those papers are of high quality or not. Therefore, it is possible that the model could generate sentences of higher quality than those written by humans. However, since we did not perform human evaluation, we pseudo-evaluated the ability of the trained model to generate words for the title and abstract of the paper as a human evaluation hypothetically. If the ROUGE-TITLE or ROUGE-ABSTRACT of the sentence generated by the model is equall or higher than that of the sentence written by a human, it can be assumed that the model was able to generate a higher quality sentence. The example of a good sentences which has a higher ROUGE-TITLE and ROUGE-ABSTRACT than that of written by human is shown in Table 6.1, Table 6.2. The example of a bad sentence which has a lower ROUGE-TITLE and ROUGE-ABSTRACT than that of written by human is shown in Table 6.3, Table 6.4.

Table 6.1: Example of a good objective sentence

human-written	in this work we examined how the spatial and temporal aspects of neural responses are altered when target stimulus in electrophysiology experiments is delivered in several non overlapping distractor target odor sequences .
model-generated	in this study we used the odor fingerprinting approach to investigate the neural basis of the solitary odor discrimination in the presence of odorant stimuli .
title	dynamic contrast enhancement and flexible odor codes
abstract	sensory stimuli evoke spiking activities patterned across neurons that can decode information distributed in flexible subsets of neurons lns match results from behavioral experiments . in sum our results suggest that trade off between stability and flexibility in sensory coding can be achieved using simple computational logic .

Table 6.2: Example of a good citation sentence

human-written	on the other hand very low ribosomal density may lead to high degradation rate of messenger rna molecules .
model-generated	the most common approach to identify the molecular mechanisms of the protein coding genes involved in the regulation of protein synthesis .
title	building better drugs developing and regulating engineered therapeutic proteins .
abstract	most native proteins do not make optimal drugs and thus second and third generation of therapeutic proteins which have been engineered to improve product attributes or to enhance process characteristics are rapidly becoming the norm . there has been unprecedented progress during the past decade in the development of platform technologies that further these ends . although the advantages of engineered therapeutic proteins are considerable the alterations can affect the safety and efficacy of the drugs . we discuss both the key technological innovations with respect to engineered therapeutic proteins and advancements in the underlying basic science . the latter would permit the design of science based criteria for the prediction and assessment of potential risks and the development of appropriate risk management plans . this in turn holds promise for more predictable criteria for the licensure of class of products that are extremely challenging to develop but represent an increasingly important component of modern medical practice .

Table 6.3: Example of a bad objective sentence

human-written	in this study we demonstrate that hdac inhibition induces 0 lo mrna expression which is concomitantly associated with h3k4 trimethylation of the alox5 promoter by the mll protein .
model-generated	here we show that hdac5 and apurinic isoform 0 is required for the activation of hdac5 and that hdac5 is required for the induction of hdac5 .
title	inhibition of class hdacs abrogates the dominant effect of mll af4 by activation of wild type mll
abstract	the alox5 gene encodes 0 lipoxygenase key enzyme of inflammatory reactions which is transcriptionally activated by trichostatin . physiologically 0 lo expression is induced by calcitriol and or transforming growth factor . regulation of 0 lo mrna involves promoter activation and elongation control within the 0 portion of the alox5 gene . here we focused on the alox5 promoter region . transcriptional initiation was associated with an increase in histone h3 lysine 0 trimethylation in trichostatin inducible manner . therefore we investigated the effects of the mll protein and its derivatives mll af4 and af4 mll respectively . mll af4 was able to enhance alox5 promoter activity by 0 fold which was further stimulated when either vitamin receptor and retinoid receptor or smad3 smad4 were co transfected . in addition we investigated several histone deacetylase inhibitors in combination with gene knockdown experiments . we were able to demonstrate that combined inhibition of hdac1 0 induces alox5 promoter activity in an mll dependent manner . surprisingly constitutive activation of alox5 by mll af4 was inhibited by class hdac inhibitors by relieving inhibitory functions deriving from mll.conversely knockdown of mll increased the effects mediated by mll af4 . thus histone deacetylase inhibitors treatment seems to switch inactive mll into active mll and overwrites the dominant functions deriving from mll af4 .

Table 6.4: Example of a bad citation sentence

human-written	the few studies on normal gait patterns in the elderly have been conducted in healthy elderly people without comorbidity and either in laboratory settings or in large clinical facilities .
model-generated	in addition to the aforementioned factors such as the risk factor is also associated with the development of psychiatric disorders .
title	basic gait parameters reference data for normal subjects 0 years of age .
abstract	basic gait parameters were extracted from 0 healthy subjects 0 men and 0 women 0 to 0 years of age . the measurements were made in gait laboratory on 0 walkway . the results are presented in series of reference tables for slow normal and fast gait . mean standard deviation coefficient of variation 0 confidence intervals and 0 prediction intervals were calculated . significant sex differences exist in all gait parameters . in two way analysis of variance model there was statistically significant age variability for gait speed and step length at normal and fast gait but not for step frequency . in the step length parameter there was significant interaction effect of age and sex at normal and fast gait . the reference data are considered valid in an indoor laboratory situation .

6.2 Objective Sentence Generation Task

In terms of the average number of words in the objective sentence, the objective sentence generated by the model is shorter than that of the objective sentence written by a human. There is no significant difference in the average number of words in the generated objective sentences between models trained on different datasets.

For ROUGE-1(filtered), ROUGE-1(unfiltered), ROUGE-2, and ROUGE-L, the model trained with High Similarity Dataset showed the best performance. Since High Similarity Dataset contains sentences that are semantically similar to the objective sentence, it makes sense that it shows superior results. On the other hand, the model that performed the worst was the one trained on High Average Probability of Word Occurrenct Dataset. It is possible that nouns, adjectives, and adverbs that occur frequently in the paper are unlikely to be included in the objective sentence. Surprisingly, the models trained with Random Dataset did not perform the worst. There is a possibility that the randomly selected sentence had a high cosine similarity to the objective sentence accidentally.

In the BERTScore results, there is no significant difference between the models trained on the four different data sets. Since BERTScore is reported to be highly correlated with human ratings, it is thought that sentences that are close to the objective sentence in terms of semantics are generated.

In ROUGE-TITLE, the human-written sentences are higher than the model-generated sentences in F1, Precsion, and Recall. However, the percentage of sentences generated by the model that had a higher or equall F1 of ROUGE-TITLE than those written by humans was 17.5%. Thus, we proved that it is possible to generate objective sentences that reflect the keywords of the study better than the objective sentences written by humans although it is a very low percentage.

In ROUGE-ABSTRACT, the human-written sentences are better than model-generated sentences in F1, Precision and Recall. However, the percentage of sentences generated by the model that had a higher or equall F1 of ROUGE-ABSTRACT than those written by humans was 14.0%. Thus, we proved that it is possible to generate objective sentences that reflect the words in abstrct of the study better than the objective sentences written by humans although it is a very low percentage.

Table 6.5: Average length of objective sentences (mean/standard deviation)

Dataset	Human-written	Model-generated
High Average	29.9/8.4	25.1/6.9
Random	29.9/8.4	25.0/6.9
High Score	29.9/8.4	24.9/6.4
High Similarity	29.9/8.4	24.8/6.6

Table 6.6: Results of ROUGE-1, ROUGE-2, ROUGE-L on objective sentence generation task (mean/standard deviation)

Dataset	ROUGE-1(filtered)	ROUGE-1(unfiltered)	ROUGE-2	ROUGE-L
High Average	0.114/0.111	0.275/0.098	0.067/0.072	0.226/0.087
Random	0.136/0.119	0.285/0.099	0.071/0.074	0.233/0.088
High Score	0.156/0.128	0.296/0.103	0.078/0.079	0.241/0.092
High Similarity	0.186/0.134	0.313/0.106	0.088/0.084	0.254/0.096

Table 6.7: Results of BERTScore on objective sentence generation task (mean/standard deviation)

Dataset	BERTScore
High Average	0.858/0.022
Random	0.860/0.022
High Score	0.862/0.024
High Similarity	0.865/0.024

Table 6.8: Results of Final score on objective sentence generation task (mean/standard deviation)

Dataset	score
High Average	29.1/ 9.9
Random	31.5/10.8
High Score	29.6/10.1
High Similarity	32.3/10.8

Table 6.9: Results of ROUGE-TITLE on objective sentence generaton task (mean/standard deviation)

	Human-written			Model-generated		
Dataset	F1	Precision	Recall	F1	Precison	Recall
High Average	0.344/0.171	0.302/0.169	0.440/0.223	0.122/0.122	0.128/0.135	0.127/0.130
Random	0.344/0.171	0.302/0.169	0.440/0.223	0.139/0.130	0.145/0.143	0.146/0.139
High Score	0.344/0.171	0.302/0.169	0.440/0.223	0.157/0.136	0.163/0.151	0.165/0.147
High Similarity	0.344/0.171	0.302/0.169	0.440/0.223	0.170/0.139	0.177/0.154	0.180/0.152

Table 6.10: Ratio of sentences generated from models with higher or equall ROUGE-TITLE on objective sentence generation task

Dataset	F1	Precision	Recall
High Average	0.120	0.171	0.122
Random	0.138	0.194	0.141
High Score	0.162	0.230	0.163
High Similarity	0.175	0.247	0.176

Table 6.11: Results of ROUGE-ABSTRACT on objective sentence generaton task (mean/standard deviation)

	Human-written			Model-generated		
Dataset	F1	Precision	Recall	F1	Precison	Recall
High Average	0.174/0.072	0.656/0.187	0.103/0.049	0.068/0.047	0.362/0.225	0.038/0.028
Random	0.174/0.072	0.656/0.187	0.103/0.049	0.076/0.049	0.398/0.231	0.043/0.029
High Score	0.174/0.072	0.656/0.187	0.103/0.049	0.083/0.054	0.435/0.240	0.047/0.032
High Similarity	0.174/0.072	0.656/0.187	0.103/0.049	0.089/0.054	0.462/0.237	0.050/0.033

Table 6.12: Ratio of sentences generated from models with higher or equall ROUGE-ABSTRACT on objective sentenge generation task

Dataset	F1	Precision	Recall
High Average	0.089	0.156	0.105
Random	0.104	0.181	0.124
High Score	0.131	0.221	0.155
High Similarity	0.140	0.244	0.163

6.3 Citation Sentence Generation Task

For the average number of words in a citation sentence, the sentence generated by the trained model was shorter than the sentence written by a human. For ROUGE-1(filtered), ROUGE-1(unfiltered), ROUGE-2, and ROUGE-L, BERTScore, the model trained on a specific dataset does not show the highest performance, but for ROUGE-1(filtered), the model trained on High Similarity Dataset shows the highest results. However, there was a significant difference in Final Score results. The model trained on Hish Similarity Dataset gave the best results, and this model can be used to generate better citation sentences. The model trained with High Average Probability of Word Occurrence Dataset had the second highest Final Score. Therefore, sentences containing words with a high probability of occurrence are considered to be effective in generating citation sentences.

For ROUGE-TITLE, human-written citation sentences performed better than model-generated citation sentences for F1, Recall, and Precision. However, 9.0% of the citation sentences generated by the model had a higher or equall ROUGE-TITLE than the citation sentences written by humans. Similarly, for ROUGE-ABSTRACT, the human-written citation sentences are better than the model-generated citation sentences for F1, Recall, and

Precision. However, 11.8% of the citatoin sentences generated by the model had a higher or equall ROUGE-ABSTRACT than the citation sentences written by humans. Thus, we have proved that, although it is a very small percentage, the system is able to generate citation sentences that reflect the important keywords of the cited documents better than humans in the citation generation task.

Table 6.13: Average length of citation sentences (mean/standard deviation)

Dataset	Human-written	Model-generated
High Average	27.2/7.9	22.5/5.6
Random	27.2/7.9	22.1/5.4
High Score	27.2/7.9	22.7/5.7
High Similarity	27.2/7.9	22.9/7.3

Table 6.14: Results of ROUGE-1, ROUGE-2, ROUGE-L on itation sentence generation task (mean/standard deviation)

Dataset	ROUGE-1(filtered)	ROUGE-1(unfiltered)	ROUGE-2	ROUGE-L
High Average	0.030/0.057	0.151/0.077	0.013/0.028	0.127/0.063
Random	0.029/0.057	0.153/0.077	0.013/0.029	0.128/0.062
High Score	0.036/0.064	0.159/0.081	0.015/0.032	0.132/0.064
High Similarity	0.040/0.066	0.156/0.080	0.015/0.031	0.128/0.064

Table 6.15: Results of BERTScore on citation sentence generation task (mean/standard deviation)

Dataset	BERTScore
High Average	0.825/0.019
Random	0.828/0.018
High Score	0.830/0.018
High Similarity	0.828/0.020

Table 6.16: Results of Final score on citation sentence generation task (mean/standard deviation)

Dataset	score
High Average	20.7/7.2
Random	13.3/4.5
High Score	14.4/4.8
High Similarity	21.7/7.6

Table 6.17: Results of ROUGE-TITLE on citation sentence generaton task (mean/standard deviation)

	Human-written			Model-generated		
Dataset	F1	Precision	Recall	F1	Precison	Recall
High Average	0.174/0.137	0.145/0.124	0.256/0.209	0.025/0.056	0.024/0.054	0.031/0.074
Random	0.174/0.137	0.145/0.124	0.256/0.209	0.024/0.056	0.023/0.055	0.030/0.072
High Score	0.174/0.137	0.145/0.124	0.256/0.209	0.032/0.065	0.031/0.064	0.039/0.085
High Similarity	0.174/0.137	0.145/0.124	0.256/0.209	0.031/0.064	0.029/0.060	0.039/0.084

Table 6.18: Ratio of sentences generated from models with higher or equall ROUGE-TITLE on citation sentenge generation task

Dataset	F1	Precision	Recall
High Average	0.074	0.083	0.084
Random	0.070	0.079	0.080
High Score	0.090	0.103	0.101
High Similarity	0.087	0.099	0.102

Table 6.19: Results of ROUGE-ABSTRACT on citation sentence generaton task (mean/standard deviation)

	Human-written			Model-generated		
Dataset	F1	Precision	Recall	F1	Precison	Recall
High Average	0.099/0.057	0.385/0.196	0.058/0.036	0.026/0.029	0.134/0.149	0.015/0.017
Random	0.099/0.057	0.385/0.196	0.058/0.036	0.027/0.030	0.141/0.149	0.015/0.017
High Score	0.099/0.057	0.385/0.196	0.058/0.036	0.029/0.030	0.152/0.154	0.016/0.018
High Similarity	0.099/0.057	0.385/0.196	0.058/0.036	0.031/0.033	0.153/0.155	0.018/0.019

Table 6.20: Ratio of sentences generated from models with higher or equall ROUGE-ABSTRACT on citation sentenge generation task

Dataset	F1	Precision	Recall
High Average	0.097	0.134	0.117
Random	0.104	0.143	0.124
High Score	0.109	0.156	0.129
High Similarity	0.118	0.156	0.142

6.4 Additional Experiement Results

We will conduct additional experiments by adapting test data of High Average Probability of Word Occurrence Dataset to the model trained on High Similarity Dataset. The models trained on High Similarity Dataset performed best in both the objective sentence generation task and the citation sentence generation task. However, when building High Similarity Dataset, we needed vectors of the objective sentence or the citation sentence in order to extract sentences with high cosine similarity to those sentences. When a human wants to generate the objective sentence or the citation sentence after writing Method, Results, Discussion and similar sections, the vector of those sentences cannot be obtained because the objective and citation sentences have not been written yet. Therefore, it is not possible to perform the same procedures as those used to build High Similarity Dataset. However, it is possible to calculate the probability of a a word's occurrence. Hence, it is necessary to verify whether High Average Probability of Word Occurrence Dataset can be used to generate high quality sentences from a model trained on High Similarity Dataset.

6.4.1 Objective Sentence Generation Task

First, in the objective sentence generation task, the average number of words in the generated objective sentences was the same as before the additional experiments. Futher, the results of ROUGE-1(filtered), ROUGE-1(unfiltered), ROUGE-2, ROUGE-L, Final Score, and ROUGE-TITLE, ROUGE-ABSTRACT were better than before the additional experiment. The improvement of Final Score in the additional experiments and the increase in the percentage of generated sentences with higher or equall ROUGE-TITLE and ROUGE-ABSTRACT than the human-written target sentences indicate that the model trained on High Similarity Dataset can be used in a realistic situation where humans automatically generate objective sentences while writing Method, Resutls, and Discussion.

Table 6.21: Comparison of average length of objective sentences (mean/standard deviation)

	Human-written	Model-generated
Before	29.9/8.4	25.1/6.9
After	29.9/8.4	25.1/7.5

Table 6.22: Comparison of results of ROUGE-1, ROUGE-2, ROUGE-L on objective sentence generation task (mean/standard deviation)

	ROUGE-1(filtered)	ROUGE-1(unfiltered)	ROUGE-2	ROUGE-L
Before	0.114/0.111	0.275/0.098	0.067/0.072	0.226/0.087
After	0.141/0.125	0.277/0.101	0.070/0.074	0.227/0.090

Table 6.23: Comparison of results of BERTScore on objective sentence generation task (mean/standard deviation)

	BERTScore
Before	0.858/0.022
After	0.854/0.025

Table 6.24: Comparison of results of Final score of objective sentence generation task (mean/standard deviation)

	score
Before	29.1/ 9.9
After	31.5/10.7

Table 6.25: Comparison of results of ROUGE-TITLE on objective sentence generaton task (mean/standard deviation)

	Human-written			Model-generated		
	F1	Precision	Recall	F1	Precison	Recall
Before	0.344/0.171	0.302/0.169	0.440/0.223	0.122/0.122	0.128/0.135	0.127/0.130
After	0.344/0.171	0.302/0.169	0.440/0.223	0.146/0.132	0.153/0.146	0.155/0.143

Table 6.26: Comparison of ratio of sentences generated from models with higher or equall ROUGE-TITLE on objective sentenge generation task

	F1	Precision	Recall
Before	0.120	0.171	0.122
After	0.149	0.213	0.153

Table 6.27: Comparison of results of ROUGE-ABSTRACT on objective sentence generaton task (mean/standard deviation)

	Human-written			Model-generated		
	F1	Precision	Recall	F1	Precison	Recall
Before	0.174/0.072	0.656/0.187	0.103/0.049	0.068/0.047	0.362/0.225	0.038/0.028
After	0.174/0.072	0.656/0.187	0.103/0.049	0.088/0.059	0.453/0.251	0.050/0.036

Table 6.28: Comparison of ratio of sentences generated from models with higher or equall ROUGE-ABSTRACT on objective sentenge generation task

	F1	Precision	Recall
Before	0.089	0.156	0.105
After	0.153	0.255	0.174

6.4.2 Citation Sentence Generation Task

In the case of the citation generation task, the average number of words in the citation sentences generated by the model increased compared to before the additional experiment, but there was no significant difference. ROUGE-1(filtered) and Final Score, ROUGE-TITLE F1 and Recall, and ROUGE-ABSTRACT only showed improved results. Surprisingly, Final Score was better than that of testing High Similarity Dataset on the model trained with High Similarity Dataset. There was no change in the percentage of sentences where the ROUGE-TITLE of the model-generated citations was higher or equall than the ROUGE-TITLE of the human-written citations. However, through additional experiments, the model succeeded in generating more citations with a higher or equall ROUGE-ABSTRACT than human-written citations. Therefore, we found that the citation sentence generation task, similarly to the objective sentence generation task, can be used in a realistic situation where a human can automatically generate the citations in the introduction while writing the Method, Resutls, and Discussion.

Table 6.29: Comparison of average length of citation sentences (mean/standard deviation)

	Human-written	Model-generated
Before	27.2/7.9	22.5/5.6
After	27.2/7.9	23.0/6.9

Table 6.30: Comparison of results of ROUGE-1, ROUGE-2, ROUGE-L on citation sentence generation task (mean/standard deviation)

	ROUGE-1(filtered)	ROUGE-1(unfiltered)	ROUGE-2	ROUGE-L
Before	0.030/0.057	0.151/0.077	0.013/0.028	0.127/0.063
After	0.032/0.058	0.149/0.077	0.013/0.028	0.124/0.062

Table 6.31: Comparison of results of BERTScore on citation sentence generation task (mean/standard deviation)

	BERTScore
Before	0.825/0.019
After	0.825/0.019

Table 6.32: Comparison of results of Final score of citation sentence generation task (mean/standard deviation)

	score
Before	20.7/7.2
After	23.3/8.4

Table 6.33: Comparison of results of ROUGE-TITLE on citation generaton task (mean/standard deviation)

	Human-written			Model-generated		
	F1	Precision	Recall	F1	Precison	Recall
Before	0.174/0.137	0.145/0.124	0.256/0.209	0.025/0.056	0.024/0.054	0.031/0.074
After	0.174/0.137	0.145/0.124	0.256/0.209	0.026/0.057	0.024/0.054	0.033/0.077

Table 6.34: Comparison of ratio of sentences generated from models with higher ROUGE-TITLE on citation sentenge generation task

	F1	Precision	Recall
Before	0.074	0.083	0.084
After	0.074	0.082	0.087

Table 6.35: Comparison of results of ROUGE-ABSTRACT on citation sentence generaton task (mean/standard deviation)

	Human-written			Model-generated		
	F1	Precision	Recall	F1	Precison	Recall
Before	0.099/0.057	0.385/0.196	0.058/0.036	0.026/0.029	0.134/0.149	0.015/0.017
After	0.099/0.057	0.385/0.196	0.058/0.036	0.028/0.031	0.138/0.145	0.016/0.018

Table 6.36: Comparison of ratio of sentences generated from models with higher ROUGE-ABSTRACT on citation sentenge generation task

	F1	Precision	Recall
Before	0.097	0.134	0.117
After	0.108	0.137	0.129

Chapter 7

Conclusion

The goal of this study is the automatic generation of objective sentences and citation sentences included in introductions for academic papers. We used open access articles from Nature for our scholarly paper dataset. In the objective sentence and citation sentence generation tasks, ROUGE-1, which was limited to nouns, adjectives, and adverbs, yielded results of 0.186 and 0.040, respectively. When we compared the quality of sentences written by humans and those generated by the model, we succeeded in generating 17.5% and 11.8% higher quality sentences than humans in the objective sentence generation task and the citation sentence generation task, respectively. Therefore, although the probability is very low, it is possible to generate higher quality objective sentences and citation sentences than humans by using the method in this study.

As a future study, it is necessary to test whether the objective sentence generation model and the citation sentence generation model constructed in this study can generate helpful objective sentences and citation sentences when people use these models. A more challenging language generation task that needs to be addressed is the automatic generation of introductions. In order to achieve this, we need to analyze the characteristics of sentences other than the objective and citation sentences in the introduction in association with the Method, Results, and Discussion of the paper. We hope that this study will reduce the burden of writing papers and assist those who are not accustomed to writing papers in English.

Acknowledgement

This research was completed under the supervision of Associate Professor Tetsuya Maeshiro. Without his support, I'm sure I wouldn't have been able to accomplish this goal on my own. He not only provided me with the computational resources to process the huge amount of data that I dealt with in this research, but also gave me casual but very essential advice to make my research better.

References

[1] Demir, S., Mutlu, U., and Özdemir, Ö. (2019). Neural Academic Paper Generation. arXiv preprint arXiv:1912.01982.

[2] Wang, Q., Zhou, Z., Huang, L., Whitehead, S., Zhang, B., Ji, H., and Knight, K. (2018, July). Paper Abstract Writing through Editing Mechanism. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) (pp. 260-265).

[3] Lloret, E., Romá-Ferri, M. T., and Palomar, M. (2013). COMPENDIUM: A text summarization system for generating abstracts of research papers. Data and Knowledge Engineering, 88, 164-175.

[4] Hu, Y., and Wan, X. (2014, October). Automatic generation of related work sections in scientific papers: an optimization approach. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 1624-1633).

[5] Meng, R., Zhao, S., Han, S., He, D., Brusilovsky, P., and Chi, Y. (2017). Deep keyphrase generation. arXiv preprint arXiv:1704.06879.

[6] Xing, X., Fan, X., and Wan, X. (2020, July). Automatic Generation of Citation Texts in Scholarly Papers: A Pilot Study. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 6181-6190).

[7] Radev, D. R., Muthukrishnan, P., Qazvinian, V., and Abu-Jbara, A. (2013). The ACL anthology network corpus. Language Resources and Evaluation, 47(4), 919-944.

[8] Lo, K., Wang, L. L., Neumann, M., Kinney, R., and Weld, D. S. (2020, July). S2orc: The semantic scholar open research corpus. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 4969-4983).

[9] Schwartz, A. S., and Hearst, M. A. (2002). A simple algorithm for identifying abbreviation definitions in biomedical text. In Biocomputing 2003 (pp. 451-462).

[10] Bird, S., Klein, E., and Loper, E. (2009). Natural language processing with Python: analyzing text with the natural language toolkit. "O'Reilly Media, Inc.".

[11] Devlin, J., Chang, M. W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

[12] Loshchilov, I., and Hutter, F. (2017). Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101.

[13] Cock, P. J., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., and De Hoon, M. J. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics, 25(11), 1422-1423.

[14] Sennrich, R., Haddow, B., and Birch, A. (2015). Neural machine translation of rare words with subword units. arXiv preprint arXiv:1508.07909.

[15] Reimers, N., and Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. arXiv preprint arXiv:1908.10084.

[16] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In Advances in neural information processing systems (pp. 5998-6008).

[17] Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. (2019). Bertscore: Evaluating text generation with bert. arXiv preprint arXiv:1904.09675.

[18] Lin, C. Y. (2004, July). Rouge: A package for automatic evaluation of summaries. In Text summarization branches out (pp. 74-81).

[19] Zaheer, M., Guruganesh, G., Dubey, A., Ainslie, J., Alberti, C., Ontanon, S., Pham, P., Ravula, A., Wang, Q., Yang, L., and Ahmed, A. (2020). Big bird: Transformers for longer sequences. arXiv preprint arXiv:2007.14062.

[20] Barua, A., Nava-Sedeño, J. M., Meyer-Hermann, M., and Hatzikirou, H. (2020). A least microenvironmental uncertainty principle (LEUP) as a generative model of collective cell migration mechanisms. Scientific reports, 10(1), 1-13.

[21] Kingma, D. P., and Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

[22] Kamali, M., and Ghahremaninezhad, A. (2018). Effect of Biomolecules on the Nanostructure and Nanomechanical Property of Calcium-Silicate-Hydrate. Scientific Reports, 8(1), 1-16.

[23] Talukder, A. K., Rashid, M. B., Yousef, M. S., Kusama, K., Shimizu, T., Shimada, M., Suarez, S. S., Imakawa, K., and Miyamoto, A. (2018). Oviduct epithelium induces interferon-tau in bovine Day-4 embryos, which generates an anti-inflammatory response in immune cells. Scientific reports, 8(1), 1-13.

[24] Lu, J., Chen, W., Huo, Y., Huang, X., and Hou, J. (2014). Clinical features and treatment outcome in newly diagnosed Chinese patients with multiple myeloma: results of a multicenter analysis. Blood cancer journal, 4(8), e239-e239.

[25] Walter, K., Cockerill, P. N., Barlow, R., Clarke, D., Hoogenkamp, M., Follows, G. A., Richards, S. J., Cullen, M. J., Bonifer, C., and Tagoh, H. (2010). Aberrant expression of CD19 in AML with t (8; 21) involves a poised chromatin structure and PAX5. Oncogene, 29(20), 2927-2937.

[26] Mani, I. (2001). Automatic summarization (Vol. 3). John Benjamins Publishing.

[27] Duan, N., Tang, D., Chen, P., and Zhou, M. (2017, September). Question generation for question answering. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (pp. 866-874).

[28] Draheim, K. M., Chen, H. B., Tao, Qs., Moore, N., Roche, M., and Lyle, S. (2010). ARRDC3 suppresses breast cancer progression by negatively regulating integrin β 4. Oncogene, 29(36), 5032-5047.

[29] Ware, M., and Mabe, M. (2015). The STM report: An overview of scientific and scholarly journal publishing.

[30] Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. Advances in neural information processing systems, 27, 3104-3112.