

A corpus for Automatic Article Analysis

Elena Callegari^{1,2}, Desara Xhura²

¹University of Iceland, Language & Technology Lab, Árnagarði við Suðurgötu, IS-101 Reykjavík

²SageWrite ehf., Miðbær, IS-101 Reykjavík

Abstract

We describe the structure and creation of the SageWrite corpus. This is a manually annotated corpus created to support automatic language generation and automatic quality assessment of academic articles. The corpus currently contains annotations for 100 excerpts taken from various scientific articles. For each of these excerpts, the corpus contains (i) a draft version of the excerpt (ii) annotations that reflect the stylistic and linguistics merits of the excerpt, such as whether or not the text is clearly structured. The SageWrite corpus is the first corpus for the fine-tuning of text-generation algorithms that specifically addresses academic writing.

Keywords


Natural Language Generation, Automatic quality assessment of text, Scientific articles, Academic writing,

1. Introduction

The latest developments in Natural Language Processing (NLP) and Natural Language Generation (NLG) demonstrate a significant gain in performance on many domain-specific NLP tasks, by pre-training on a large corpus of text and fine-tuning using prompt engineering¹ in specific task [1][2][3]. The SageWrite corpus is a manually annotated corpus created as a training dataset for the development of automatic text-generation and quality-assessment tools for academic writing².

When writing the different sections of an academic paper, authors often start by creating a rough draft or outline of what they want that section to say, which they then proceed to edit -and re-edit- until they are satisfied with it. An author writing the introduction of a linguistics paper may for example start by writing something along the lines of i, which they will then proceed to edit until it looks something like ii:

- i. *My intentions:*
 - first: present core data on focus particles*
 - second, review different existing approaches*
 - 3rd: say what I think about what works best*

 ecallegari@hi.is (E. Callegari); dxhura@gmail.com (D. Xhura)



© 2022 Copyright for this paper by its Authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

¹Prompt engineering is a way of fine-tuning, where the NLP algorithm gets fed with examples of input and expected results.

²In the future, the dataset could also be relevant for text summarization purposes similar to [4]

- ii. My intentions in this article are threefold: first, to outline the key data that any successful account of focus particles should explain; second, to review existing approaches that attempt to account for these data; and third, to offer my own views about the direction any successful analysis should take.

Our primary goal is automate the process that leads from i to ii: we want to generate grammatical text starting from a rough draft of what the final text should look like. Put differently, what we aim to do is to streamline the revision process that leads from i to ii. What is required to generate ii out of i stands halfway between natural language generation out of a limited input [5] and advanced automatic paraphrasing[6]. Our secondary goal is to develop a classifier that can process scientific articles and automatically assess whether or not they exhibit certain qualities or flaws that we deem relevant to assess scientific publications, such as whether or not information is clearly presented and whether or not the text exhibits a good flow. Again, this is in an attempt to streamline the revision process: if the stylistic shortcomings of a paper are flagged automatically, the author(s) of said paper can more readily address them. The SageWrite corpus was created to assist in the training of both of these functionalities. A first version of the corpus (version 0.1), consisting of 100 annotated excerpts, was published online in February 2022³. We plan on increasing the size of this dataset as more excerpts get annotated.

2. Text Selection

The 100 manually annotated excerpts were extracted from various types of academic articles. To obtain the excerpts, we first created a database containing scientific articles taken from Arxiv, PubMed, plus around 70 articles that we randomly selected from various disciplines in the Humanities. The articles taken from Arxiv were all dated March 2020 onwards.

To extract the excerpts, we wrote a Python program that automatically extracted excerpts of around 300 words from various points in an article. This was done to ensure that text belonging to various sections of a paper (e.g. introduction, abstract, conclusions) was included. Text was always selected from the beginning of a paragraph until the end of a paragraph. The average length of the excerpts was 193 words.

The excerpts were annotated by three annotators. As we wanted to work on academic texts, we hired annotators who had ties with academia and experience with academic writing. Accordingly, one of our annotators was a MA student, one was a university lecturer and one had a PhD degree. All annotators were also native speakers of (American) English.

Annotations were completed online on a dedicated platform where annotators could automatically log each part of the annotation for a given excerpt.

Annotators saw rotations consisting of one excerpt from a PubMed article, one from an Arxiv article and one from our Humanities articles. As we thought it would be interesting to see how different individuals would react to the same text, all annotators saw and hence annotated the same excerpts.

³<https://github.com/elenaSage/SageWrite0.1corpus>

3. Structure of the Corpus

For each of the 100 excerpts, the corpus contains (A) three corresponding rough-draft versions of excerpt, each authored by a different annotator and (B) a list of tags that describe the stylistic and linguistic qualities of each excerpt.

3.1. The Drafts

When writing up a section of an academic paper, authors generally start out by writing a rough draft of what they want to say. This draft differs from the final version of a paper in several respects: a draft may contain various types of abbreviations (e.g. i), is generally more schematic in nature (e.g. articles, copulas, 1st person singular pronouns may be dropped, arrows and empty lines may be used in place of linkers and connectors (e.g. ii)), and frequently contains (overly) colloquial language ((e.g. iii).

- i. a) contractions: "that's" vs. "that is"
 b) colloquialisms: "cause"/"coz" for "because", "w" in place of "with"
 c) other types of abbreviations: "mvt" for "movement", "foc" for "(linguistic) Focus"
- ii. a) "in sect 1, will be talking about foc marking patterns in Malayalam"
 b) "in sect 1 -> foc patterns in Malayalam"
- iii. a) "In the essay Racisms, Kwame Anthony Appiah says what he thinks about the topic"
 vs.
 "In the essay Racisms, Kwame Anthony Appiah provides his thoughts on this issue."
 b) "But are MCI patients actually aware of their cognitive deficits? That's debatable "
 vs.
 "However, whether patients with MCI are truly aware of the full extent of their cognitive deficits is a matter of debate"

We asked our three annotators to read each excerpt and try to reverse-engineer what the draft version of that excerpt might have looked like, and to write that down. We asked them to experiment with different drafting styles; for example, we explained that while some authors might use lots of abbreviations, others might prefer to spell out every or most words. While some authors might use extremely colloquial language, others might prefer to adhere to academic lexical standards already in earlier versions of a paper.

As we are dealing with academic text, our goal is to develop NLG tools that do not generate too much beyond the original input: should the AI generate too much on top of the initial input provided by the user, one could question whether the resulting generated text is truly the work of the author or rather should be considered the work of the AI. Because of these concerns, we instructed our annotators not to leave out non-recoverable information from the drafts. For example, information occurring between parentheses in the original text was always included in the corresponding draft version (see 3.1).

- i. a) **Original Text**
 "It also presents methods that may be used for analyzing language interplays in

general (**demonstrated using the PDT data**)”

b) **Draft Version** (as created by Annotator 2)

”Present methods to analyze language interplays in general (**see PDT**)”

Annotators first practiced annotation on a set containing 50 sample excerpts. During this practice run, annotators got direct feedback by the authors of this paper, who reviewed the annotations of the sample excerpts. These 50 practice excerpts are not included in the dataset we published online.

3.2. The Tags

We asked our annotators to evaluate the stylistic and linguistic merits of each excerpt by selecting dedicated tags. We started out with a set of 13 tags that we came up with ourselves, based on our own personal perception of what common issues are found in scientific articles, as well as on the literature on the topic[7],[8][9][10]. The 13 initial tags are listed below; we also provide a short explanation of those tags which may not be fully transparent.

- i. Colloquial Language: to be used whenever overly colloquial language is used;
- ii. Formal Language: whenever excessively formal language is used, e.g. when expressions like *et ceteris paribus* are used (too often);
- iii. Jumbled Vocabulary: to describe combinations of words that make little sense, e.g. “the council has a *strong objective*”(objectives cannot be *strong*);
- iv. Unnecessary jargon;
- v. Verbosity;
- vi. Opaque writing: for text that is obscure, hard to understand;
- vii. Overly long sentences;
- viii. Abuse of passive sentences: e.g. “It has been found that there had been many ...”;
- ix. Excessively complex syntax: e.g. “It is expected that an exploration of the variables affecting the effectiveness of reading aloud will support us in designing lessons (...)”;
- x. Clear Structure: to mark text that is clear and well-structured, text that clearly communicates the writer’s intentions, data or results;
- xi. Pretentiousness;
- xii. Engaging Writing: text that is compelling, witty and makes one want to read more;
- xiii. Dull writing: text that is dry, boring and not engaging;

When selecting which tags to include in our inventory, we tried including tags that refer to different linguistic dimensions. For example, tags 1 to 5 relate to the **lexical** dimension, tags 7 to 10 capture **syntactic** properties, tag 10 relates to **pragmatics** and tag 11 to 13 relate to the perceived **stylistic** merits or demerits of a text. We also tried to balance the number of positive and negative tags. We provided annotators with a document explaining each tag and where it should be used, which we went over together. We then let the annotators try out the tags over the 50 sample excerpts, providing them with personalized feedback and comments should they appear to be using some of the tags incorrectly. We also told annotators that they could suggest additional tags should they notice anything that was obviously missing. After

this initial dry-run over the 50 sample excerpts, based on the suggestions from the annotators we added 6 additional tags:

- i. Redundant (content): to be used for words, phrases or clauses that are superfluous;
- ii. Repetition (style): for anaphoric repetitions, epiphoric repetitions and anytime sentence structure or vocabulary is not diverse enough; A problem which is encountered frequently in academic writing [11]
- iii. Poor flow: if the logical flow of a text is whacky, or whenever there are no clear threads to follow;
- iv. Non-sequitur: sentences that do not follow logically from anything that was said before;
- v. Unclear/vague: for unclear referents, ambiguous statements and anything that should have been explained in more detail;
- vi. Fragment: for sentences/ paragraphs that feel excessively telegraphic in style.

Also based on the suggestions from the annotators, we replaced the tag “jumbled vocabulary” with “word choice”:

- word choice: to be used for any questionable lexical choice, *whether at the sentence level or at the level of single words.*

The final tag inventory thus consisted of 19 tags. Annotators were given the option to select tags either globally or locally. Locally selected tags referred to specific sub-parts of an excerpt, e.g. to specific words, phrases, paragraphs. An example would be the tag “overly long sentences”, that could apply to a single sentence. A tag that was selected globally meant that the specific characteristic that the tag singled out applied to the entire excerpt; an example would be the tag “poor flow”. In the corpus, each excerpt is associated with each of the 19 tags, and for each excerpt each of the 19 tags has a value ranging from 0 to 3: 3 if that tag was selected for that excerpt by all three annotators, 0 if it was selected by no annotator. To simplify the structure of the corpus, we eliminated the distinction between global and local tags (in version 0.1 at least): if a tag was selected by an annotator, it is associated with a “1” value, regardless of whether the tag was selected globally or locally. The same holds for cases in which the same tag was selected locally more than once within the same excerpt. In future versions of the corpus, we plan on making the distinction between global and local tags accessible.

4. Exploratory Data Analysis

4.1. Tags used

Table 1 below illustrates how often the tags were selected at least once for a given excerpt (whether locally or globally) by an annotator. We see that the most frequently selected tags were “opaque writing” (34 instances), “clear structure” (36 instances) and “word choice” (18 instances).

Some of the tags which were relatively underused are “formal language” (1 instance), “colloquial language” (2 instances), “repetition” (2 instances) and “abuse of passive sentences” (3 instances). There are different reasons that could explain why these tags were underused: the

low frequency of "colloquial language" could be explained by assuming that academic papers displaying an overly colloquial style are fairly rare; if anything, academic papers tend to be *too* formal. The low frequency of the "formal language" tag could be explained by citing difficulties in determining when text is *too* formal in a field where the use of formal language is generally encouraged. The same explanation could be extended to account for the low frequency of "abuse of passive sentences": passive sentences are a feature of academic writing. Annotators might have felt compelled to accept as good passive structures that they would have flagged otherwise precisely because they were aware they were dealing with academic text.

colloquial language	2	abuse of passives	3	repetition	2
formal language	1	clear structure	36	fragment	9
jargon	4	pretentiousness	3	non-sequitur	4
verbosity	2	engaging	13	poor flow	8
opaque writing	34	dull	10	redundant	6
overly long sentences	4	unclear	12	complex syntax	4
word choice	18				

Table 1
Frequency of Tag Usage in corpus

4.2. Length of Drafts

Figure 1 illustrates the length distribution of each of the 100 excerpts. The average length of the excerpts was 193 words.

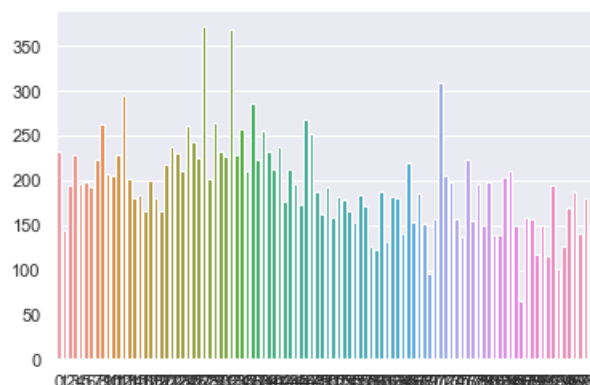


Figure 1: Length in words of each of the original 100 excerpts

Figure 2, 3 and 4 illustrate the length distribution of each of the drafts created by annotator 1, 2, and 3 respectively. The average amount of words in the drafts was 146.5 for annotator 1, 155 for annotator 2 and 119 for annotator 3.

Note that some of the data points are missing in the figure for annotators 2 and 3. This is because annotators were instructed not to annotate excerpts that would be too time-consuming

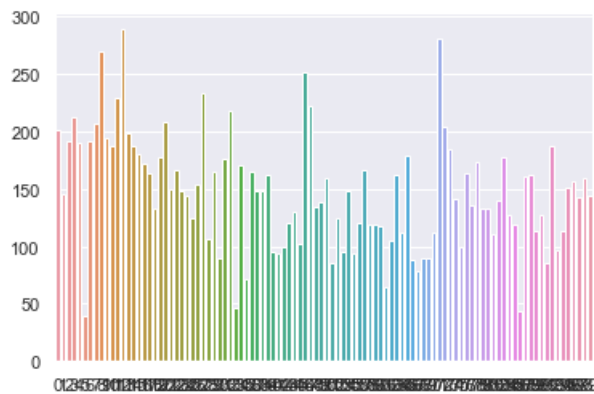


Figure 2: Length of each of the drafts created by annotator 1

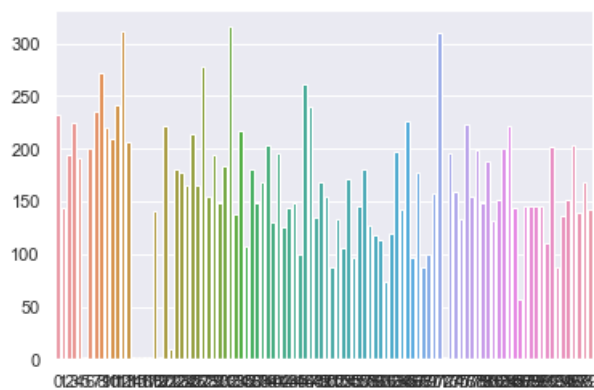


Figure 3: Length of each of the drafts created by annotator 2

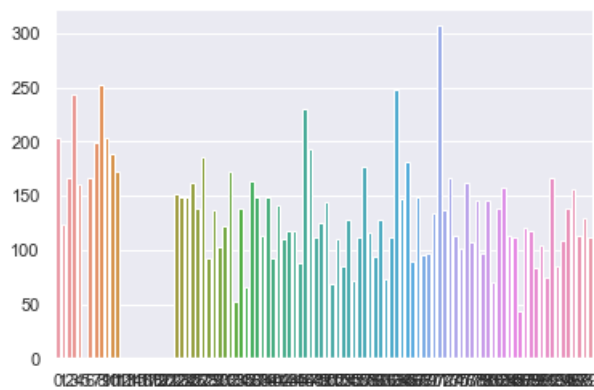


Figure 4: Length of each of the drafts created by annotator 3

to annotate, e.g. excerpts containing lots of formulas or symbols. The missing data points in Fig 3-4 then represent excerpts that the annotators decided not to annotate.

Figures 5, 6 and 7 illustrate the ratio between length of the original excerpt and the corresponding draft for each of the 3 annotators. For annotator 1, the average ratio corresponds to 0.774; for annotator 2, to 0.813; for annotator 3, to 0.633. We see that the length of a draft increases more or less incrementally with the length of the original text for annotators 1 and 2. In the case of annotator 3, on the other hand, the length of the initial outline is less reliable of an indicator of the length of the corresponding draft.

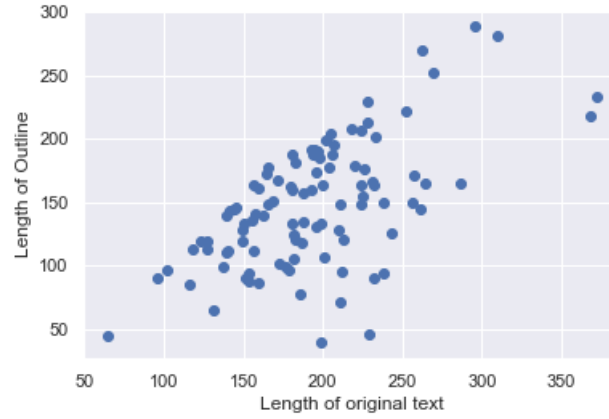


Figure 5: Ratio between length of excerpts and corresponding draft for annotator 1

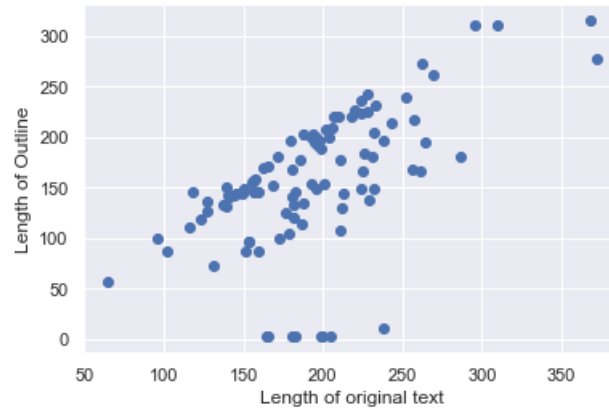


Figure 6: Ratio between length of excerpts and corresponding draft for annotator 2

References

- [1] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, G. Neubig, Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing, 2021.

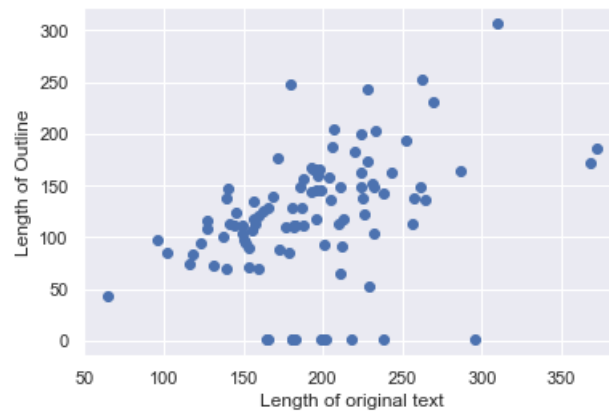


Figure 7: Ratio between length of excerpts and corresponding draft for annotator 3

arXiv:2107.13586.

- [2] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners, 2020. arXiv:2005.14165.
- [3] X. Han, Z. Zhang, N. Ding, Y. Gu, X. Liu, Y. Huo, J. Qiu, Y. Yao, A. Zhang, L. Zhang, W. Han, M. Huang, Q. Jin, Y. Lan, Y. Liu, Z. Liu, Z. Lu, X. Qiu, R. Song, J. Tang, J.-R. Wen, J. Yuan, W. X. Zhao, J. Zhu, Pre-trained models: Past, present and future, 2021. arXiv:2106.07139.
- [4] E. Collins, I. Augenstein, S. Riedel, A supervised approach to extractive summarisation of scientific papers, 2017. arXiv:1706.03946.
- [5] e. a. Qu, Yuanbin, A text generation and prediction system: pre-training on new corpora using bert and gpt-2. 2020 IEEE 10th International Conference on Electronics Information and Emergency Communication (ICEIEC) (2020).
- [6] H. Palivela, Optimization of paraphrase generation and identification using language models in natural language processing., International Journal of Information Management Data Insights 1 (2021) 100025.
- [7] S. Pinker, Why academics stink at writing., The chronicle of higher education 61 (2014).
- [8] E. Ventola, e. Anna Mauranen, Academic writing: Intercultural and textual issues 41 (1996).
- [9] G. F. Badley, Post-academic writing: Human writing for human readers., Qualitative Inquiry 25 (2019) 180–191.
- [10] P. Crompton, Hedging in academic writing: Some theoretical problems., English for specific purposes 16 (1997) 271–287.
- [11] W. Xiao, G. Carenini, Systematically exploring redundancy reduction in summarizing long documents, 2020. arXiv:2012.00052.