4B1-3

A Corpus for Studies on Scientific Writing Assistance

Ngan L.T. Nguyen Yusuke Miyao

National Institute of Informatics

Previous research on writing assistance has mostly focused on correcting spelling and grammatical errors. However, the proofreading process, which is required in professional writing, involves not only the correction of grammatical errors, but also the paraphrasing of inarticulate sentences, when necessary, to make them more fluent. This work aims at constructing a corpus to satisfy such requirements to support research towards professional writing assistance. Our corpus is a collection of scientific work written by non-native speakers that has been proofread by native English experts. A new annotation scheme, which is based on word-alignments, is then proposed that is used to capture all types of inarticulations and their corrections including both spelling/grammatical error corrections and paraphrases made by proofreaders. The resulting corpus contains 3,485 pairs of original and revised sentences, of which, 2,516 pairs contain at least one articulation.

1. Introduction

Detection and correction of misspellings and grammatical errors have been recognized as key techniques for writing assistance, and have extensively been studied in natural language processing (NLP) [Whitelaw 09, Gamon 10, Tetreault 10, Park 11]. However, correcting misspellings and grammatical errors, which can be performed by normal English native speakers, does not satisfy all the requirements of professional writing [Futagi 10]. The core of the proofreading process, in reality, is paraphrasing inarticulations, which can only be done by expert proofreaders. Considering the two paraphrased sentences (1a) and (1b) below, we can see that sentence (1b) is likely to be considered better by most people [Williams 10], although neither of them contains any misspellings or grammatical errors.

(1a) *The outsourcing of high-tech work to Asia by corporations means the loss of jobs for many middle-class American workers.*

(1b) *Many middle-class American workers are losing their jobs, because corporations are outsourcing their high-tech work to Asia.*

[Williams 10]

Although most of the existing corpora are designed to capture errors in spelling and grammar, they have not paid enough attention to paraphrasing (see Section 2.).

We constructed a corpus that we called scientific writing assistance corpus (SWA), to support research on assistance with scientific-writing that captures all types of inarticulations, including those in both mispellings/grammar and paraphrasing. We have used the term *inarticulation* and *inarticulation correction* instead of *error* and *error correction* in this paper, to include in our task the paraphrasing, which is actually not errors.

Figure 1 overviews the methodology we used to construct the corpus. Scientific work written by non-native researchers or graduate students are collected (i.e., data collection, see Section 4.), and this was then proofread by English native experts (i.e., proofreading). After that, we preprocessed the documents to convert them into a predefined format (i.e., preprocessing, see Section 4.). The documents were then ready for the process of annotation (i.e., corpus annotation, see Section 5.). Annotators with linguistic backgrounds were asked to strictly follow our annotation scheme, which had been designed to capture all types of inarticulations (i.e., annotation scheme design, see Section 3.).

Contact: Ngan L.T. Nguyen, National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo, ngan@nii.ac.jp

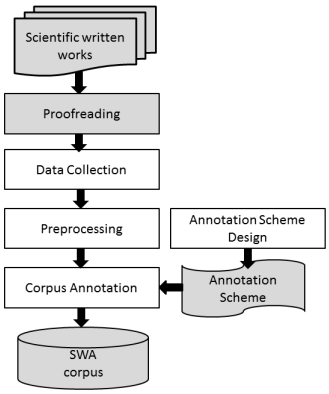

Figure 1: Methodology for corpus annotation

Our corpus construction had several substantial advantages in comparison to existing corpora: (1) the proofreading process is separated from the annotation process. By doing this, both the writers and the proofreaders were unaware of the construction of the corpus, so it could capture real articulations and corrections to these, (2) the alignment-based annotation scheme, which was originally proposed for paraphrase annotation [Cohn 08], was extended and employed in annotations of articulation correction, and (3) paraphrases were captured, and were proved to be an important type of articulation correction.

2. Alignment-based scheme for paraphrase annotation

Our annotation scheme extends the alignment-based annotation scheme for paraphrase annotation proposed by Cohn et al. [Cohn 08]. Their main idea was to use word alignments to record the correspondences in a pair of paraphrased sentences. Words or phrases that expressed the same meaning were connected via one-to-many, many-to-one, or many-to-many *bidirectional alignments* (called *bi-alignments* after this). An alignment is marked *possible* when it has a loose paraphrase relation, otherwise it is marked *certain*. Words or phrases that do not have a correspondence in

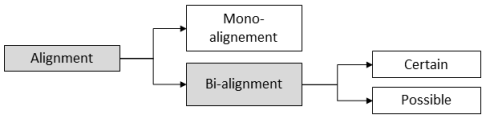

Figure 2: Illustration for Cohn et al.'s tagset. Categories in gray are used for classification but not for tagging.

the other sentence are left unaligned. These unaligned words can be considered as *mono-alignments*. Thus, there are a total of three types of annotations in their annotation scheme: mono-alignments, possible alignments, or certain alignments (Figure 2).

We observed that there were similarities between paraphrase annotation and error annotation, since the original sentence and its proofread version could be seen as a pair of sentential paraphrases. Thus, the annotation scheme by Cohn et al. [Cohn 08] fits well into our purpose of capturing paraphrases. The only difference between articulartion annotation and paraphrasing annotation is that the paraphrase relationship in articulation correction is asymmetric, which is because the proofread sentence is preferable to the original sentence. In addition, the word alignment annotations naturally captures discontinuous correspondences, which is superior to the text-span annotation employed by other learner corpora. However, as the original annotation scheme was not designed for inarticulation annotations, we needed to extend this annotation scheme by designing a taxonomy for categorizing the alignments in a way that made the annotation useful for articulation correction.

3. Annotation scheme design

3.1 Overview

We extended the alignment-based paraphrase annotation scheme of Cohn et al. [Cohn 08] by categorizing the alignments into more fine-grained types (see Figure 3) to capture all types of inarticulation corrections. Figure 4 outlines example annotations to illustrate our annotation scheme. The alignments at the top level, are divided up into four broad types: Preserved, Metadata, Inarticulation Bi-alignment and Inarticulation Mono-alignment.

The Preserved type of alignments is the most trivial type that connects words with the same surface and function, e.g., *the*, *efficiency*, *various*, *methodologies* in Figure 4(A). Still, there are many cases where two words have the same surface form, but do not have the same functions in the original and the proofread sentences. For instance, the word *of* in the above example appears in both the sentences, but the two occurrences are not aligned, because they modify different words, i.e., *approach* and *methodologies* in this case.

The tags in the Metadata group are designed to capture information that is specific to proofreading by humans. There are two tags in this group: Uncertain and Problematic. An alignment is marked as *uncertain* when the proofreader is not confident in the correction. This type is specific to the proofreading process. When the native proofreader is doubtful about his/her understanding of the original sentence, he/she will comment on it by stating *"I do not understand this,"* or *"This correction is a guess"*. An alignment is classified as *Problematic* when the annotators discover that the proofreader has made an erroneous correction. This happens when the proofreader misunderstands the author's intention. Although

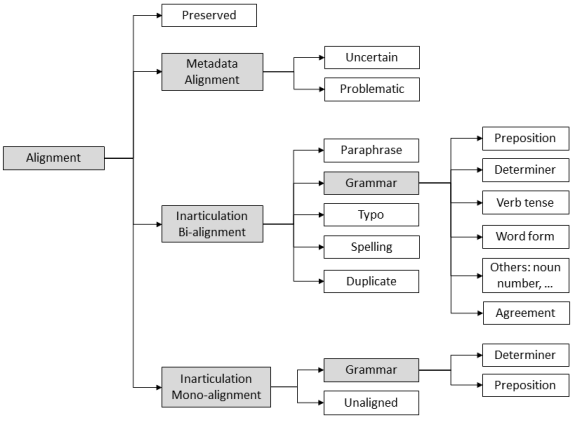

Figure 3: Proposed tagset. Categories in gray are used for classification but not for tagging.

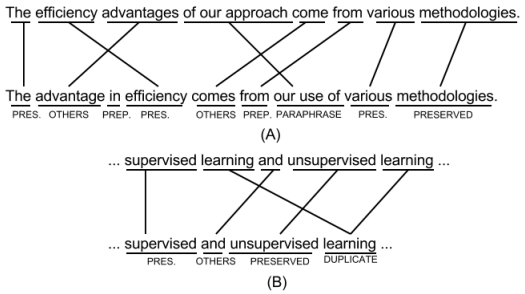

Figure 4: Examples of annotations using our annotation scheme. Top has original texts, and bottom has the proofread text.

such situations are rare, this tag is designed to offer a mechanism for annotators to provide feedback.

Inarticulation alignments including mono-alignments and bi-alignments are for capturing inarticulations and their corrections. The Grammar subtype of inarticulation alignments is not used for all types of grammatical errors as in the other annotation scheme, but is limited to some well-defined types of grammatical errors, which will be explained later in Section 3.2. The other subtypes are Duplicate, Typo, Spelling, and Unaligned, which will be explained in the following.

- **Duplicate**: A duplicate alignment connects words that appear once in the original sentence, but more than once in the proofread sentence, or vice versa. This tag captures the correction for articulations like the word *learning* in the example in Figure 4(B).

- **Spelling**: A spelling alignment is used for misspellings, e.g., *occured→occurred*[*1]. This also includes the use of hyphens, e.g., *state of the art→state-of-the-art*.

- **Typo**: The expression typo is a short form of typographical error, which refers to errors caused by typing mistakes. If annotators judge that the error is likely to be caused by a typing mistake, they should mark the errors as typo. Typo may be considered to be less important in writing assistance.

∗1 The expression to the right of the arrow (→) is the preferred expression within context of writing

- **Unaligned**: An unaligned mono-alignment is used for words in the original sentence that have no correspondences in the proofread sentence, or vice versa.

Reordering of words are naturally captured by cross alignments, so we do not create a type for this. Punctuation marks are not annotated.

3.2 Grammar

Grammar-typed alignments connect a grammatical error in the original sentence with its correction in the proofread sentence. Grammatical errors in our annotation scheme are comprised of errors with determiners, prepositions, verb tenses, word forms, agreement, and others. They are tagged with the corresponding tags called Determiner, Preposition, Verb tense, Word form, Agreement, and Others. The Others type merges several specific subtypes of grammatical errors, including noun number, verb number, wh-word choice, or conjunction choice. Note that we do not use Others as a catch-all type. Except for Agreement, most of the subtypes of the Grammar type can be aligned well with the error types in the error taxonomies used by the existing corpora. The Agreement type is used to capture the number agreements of articles and nouns, genitives and nouns, or nouns and verbs, when a change in the number of one word forces us to change the number and form of another word.

3.3 Paraphrase

Any type of correspondence that cannot be classified into these types above is marked Paraphrase. In other words, Paraphrase is used as a catch-all type. Those errors that require complex corrections, i.e., corrections to phrase structures or sentence structures, which are not classified into the Grammar type, are captured with Paraphrase. We have followed the definition of paraphrases in the guidelines for paraphrase annotation by Callison-Burch et al. [Callison-Burch 06]: *"paraphrases convey the same meaning but are worded differently"*. We have two rules of thumb for the boundary of paraphrases: (1) shorter paraphrases are preferable (similar to [Callison-Burch 06]), and (2) a paraphrase alignment should not contain an alignment of other types in it.

4. Data collection and preprocessing

We collected eighteen scientific works that were written by seven authors with two language backgrounds: Japanese and Vietnamese. The collected documents included different types of scientific publications such as short papers, full papers, and book chapters. We will use the terminology *document* to refer to a written work of any type. The collected documents belonged to two domains or fields of studies, which were computer vision (11 documents) and natural language processing (7 documents); and all were proofread by native English experts.

We then preprocessed these documents to convert them into a standard format. Non-text information such as figures and tables were removed. Format tags such as LaTeX's tags were also removed. We separated the original text and the proofread text for each document, and aligned the sentences in these two texts, so that a line in the original text corresponded to a line in the proofread text. We found that there were cases where a sentence in the original text should have been aligned with more than one sentence in the proofread text or vice versa. We allowed two or more sentences to be aligned in such cases.

	Total
Documents	18
Pairs of sentences	3,485
Pairs of sentences containing articulations	2,516
Words in original texts	75,968
Words in original texts annotated as Preserved	69,738
Inarticulation alignments	4,686
Metadata alignments	26

Table 1: Summary of statistics for SWA

Alignment Type	Count	Ratio (%)
Paraphrase	1,372	29.3
Bi-Grammar	1,511	32.2
Typo	68	1.5
Spelling	308	6.6
Duplicate	13	0.3
Preserved	2	0.0
Mono-Grammar	1,212	25.9
Unaligned	200	4.3
TOTAL	4,686	100.0

Table 2: Statistics for all alignments (except for the Preserved type) annotated in the corpus

5. Corpus annotation and results

5.1 Corpus statistics

We made use of Yawat, a web-based word-alignment annotation tool [Germann 08] to annotate the corpus. Yawat accepts text files containing pairs of aligned sentences as input. We applied a simple string-matching algorithm to produce default Preserved and Unaligned alignments for the corpus to save annotation time and effort.

The statistics for the annotated corpus are summarized in Table 1. A total of 4,686 Inarticulation alignments and 26 Metadata alignments were annotated for 2,516 pairs of sentences in 18 documents. 69,738 (91.8%) of the total of 75,968 words in the corpus were annotated with Preserved alignments.

Table 2 lists the ratios (%) of broad types of alignments. We can see that the Grammar errors, both in bi- and mono-alignments, occupy 58.1% of the total errors, which is not a surprise. Paraphrase alignments occupy a significant part, i.e., 29.3% of the total. These figures indicates that paraphrasing is an essential type for scientific writing; therefore, research on writing assistance should pay more attention to error correction by using paraphrasing.

The ratios of the subtypes of Grammar alignments are listed in the column named *SWA* (the name of our corpus) in Table 3. Out of all grammatical errors, determiners caused a lot of troubles for non-native writers from the Japanese and Vietnamese language backgrounds, even though the authors of the collected documents all had an advanced level of proficiency in English. This may be because of the difference between the characteristics of their background languages and the English language.

6. Potential use of SWA corpus

One of the immediate applications of the SWA corpus is the automatic tagging of proofreading results. Usually, proofread documents contain only inarticulation corrections and not their intentions or types, so it is useful for non-native speakers if computer systems can automatically classify the corrections. Non-native

Type	SWA Count	SWA (%)
Determiner	1,176	25.1
Preposition	547	11.7
Others	427	9.1
Verb tense	369	7.9
Word form	151	3.2
Agreement	53	1.1
TOTAL (Grammar only)	2,723	58.1
Total of all inarticulations	4,686	100.0

Table 3: Statistics for Grammar alignments in SWA in comparison with KJ corpus and NUCLE corpus with α = TOTAL$_{SWA}$/ TOTAL$_{KJ}$, β = TOTAL$_{SWA}$/ TOTAL$_{NUCLE}$

writers can query this structured corpus to learn from their own's or other people's mistakes.

The corpus can currently be used to provide benchmark data for testing the performance of NLP techniques for assistance with professional writing, including the grammatical error correction and paraphrasing techniques. Exploring the use of domain-specific knowledge and discourse information is a promising direction to improve these.

7. Conclusion

We described the SWA corpus, which was constructed to support studies on automatic writing assistance, particularly for scientific writing. The traditional problem of error annotation was viewed as a paraphrase annotation of pairs of the original and proofread sentences. This view inspired us to extend the alignment-based annotation scheme, previously used for paraphrase annotation, to our annotation process.

The SWA corpus is available for research on request basis.

References

[Callison-Burch 06] Callison-Burch, C., Cohn, T., and Lapata, M.: Annotation guidelines for paraphrase alignment (2006)

[Cohn 08] Cohn, T., Callison-Burch, C., and Lapata, M.: Constructing corpora for the development and evaluation of paraphrase systems, *Comput. Linguist.*, Vol. 34, No. 4, pp. 597–614 (2008)

[Dahlmeier 11] Dahlmeier, D. and Ng, H. T.: Grammatical error correction with alternating structure optimization, in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pp. 915–923, Stroudsburg, PA, USA (2011), Association for Computational Linguistics

[Dale 10] Dale, R. and Kilgarriff, A.: Helping Our Own: The HOO 2011 Pilot Shared Task, in *Proceedings of the 13th European Workshop on Natural Language Generation*, Dublin, Ireland (2010)

[Dale 11] Dale, R. and Kilgarriff, A.: Helping Our Own: Text Massaging for Computational Linguistics as a New Shared Task, in *Proceedings of the International Natural Language Generation Conference 2011*, Nancy, France (2011)

[Futagi 10] Futagi, Y.: The effects of learner errors on the development of a collocation detection tool, in *Proceedings of the fourth workshop on Analytics for noisy unstructured text data*, AND '10, pp. 27–34, New York, NY, USA (2010), ACM

[Gamon 10] Gamon, M.: Using mostly native data to correct errors in learners' writing: a meta-classifier approach, in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pp. 163–171, Stroudsburg, PA, USA (2010), Association for Computational Linguistics

[Germann 08] Germann, U.: Yawat: yet another word alignment tool, in *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Demo Session*, HLT-Demonstrations '08, pp. 20–23, Stroudsburg, PA, USA (2008), Association for Computational Linguistics

[Izumi 04] Izumi, E., Uchimoto, K., and Isahara, H.: SST speech corpus of Japanese learners' English and automatic detection of learners' errors, Vol. 28, pp. 31–48 (2004)

[Milton 10] Milton, J. and Cheng, V. S. Y.: A toolkit to assist L2 learners become independent writers, in *Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics and Writing: Writing Processes and Authoring Aids*, CL&W '10, pp. 33–41, Stroudsburg, PA, USA (2010), Association for Computational Linguistics

[Nagata 11] Nagata, R., Whittaker, E., and Sheinman, V.: Creating a manually error-tagged and shallow-parsed learner corpus, in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pp. 1210–1219, Stroudsburg, PA, USA (2011), Association for Computational Linguistics

[Park 11] Park, Y. A. and Levy, R.: Automated whole sentence grammar correction using a noisy channel model, in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pp. 934–944, Stroudsburg, PA, USA (2011), Association for Computational Linguistics

[Stein 10] Stein, B., Potthast, M., and Trenkmann, M.: Retrieving customary web language to assist writers, in *Proceedings of the 32nd European conference on Advances in Information Retrieval*, ECIR'2010, pp. 631–635, Berlin, Heidelberg (2010), Springer-Verlag

[Tetreault 10] Tetreault, J., Foster, J., and Chodorow, M.: Using Parse Features for Preposition Selection and Error Detection, in *Proceedings of the ACL 2010 Conference Short Papers* (2010)

[Whitelaw 09] Whitelaw, C., Hutchinson, B., Chung, G. Y., and Ellis, G.: Using the web for language independent spellchecking and autocorrection, in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2 - Volume 2*, EMNLP '09, pp. 890–899, Stroudsburg, PA, USA (2009), Association for Computational Linguistics

[Williams 10] Williams, J. M. and Colomb, G. G.: *Style: Lessons in clarity and grace*, Boston, MA: Longman (2010)