Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 4663–4673
Marseille, 20-25 June 2022

Automating Idea Unit Segmentation and Alignment for Assessing Reading Comprehension via Summary Protocol Analysis

Marcello Gecchele†, Hiroaki Yamada†, Takenobu Tokunaga†
Yasuyo Sawaki‡, Mika Ishizuka§
† Tokyo Institute of Technology, ‡ Waseda University, § Tokyo University of Technology
Tokyo, Japan
gecchele.m.aa@m.titech.ac.jp, yamada@c.titech.ac.jp, take@c.titech.ac.jp,
ysawaki@waseda.jp, ishizuka@stf.teu.ac.jp

Abstract

In this paper, we approach summary evaluation from an applied linguistics (AL) point of view. We provide computational tools to AL researchers to simplify the process of Idea Unit (IU) segmentation. The IU is a segmentation unit that can identify chunks of information. These chunks can be compared across documents to measure the content overlap between a summary and its source text. We propose a full revision of the annotation guidelines to allow machine implementation. The new guideline also improves the inter-annotator agreement, rising from 0.547 to 0.785 (Cohen's "κ"). We release L2WS 2021, a IU gold standard corpus composed of 40 manually annotated student summaries. We propose IUExtract; i.e. the first automatic segmentation algorithm based on the IU. The algorithm was tested over the L2WS 2021 corpus. Our results are promising, achieving a precision of 0.789 and a recall of 0.844. We tested an existing approach to IU alignment via word embeddings with the state of the art model SBERT. The recorded precision for the top 1 aligned pair of IUs was 0.375. We deemed this result insufficient for effective automatic alignment. We propose "SAT", an online tool to facilitate the collection of alignment gold standards for future training.

Keywords: Idea Unit, Student Summary Evaluation, Summary Response Analysis, Segmentation, Alignment

1. Introduction

In second language learning, summaries are among the most popular type of student assignments. Summaries are effective tools to teach techniques for the identification and concise delivery of information (Graham and Perin, 2007), to assess the language comprehension ability of students (Westby et al., 2010) and to teach paraphrasing skills; i.e., the ability to rewrite information adopting a different expression. Such techniques are paramount for academic writing as they are necessary to avoid plagiarism (Keck, 2014). While writing a summary, students are asked to read a source text, pick the core information and rewrite it in a shorter text.

In this paper, we are looking to evaluate English language summaries in regards to their content, assessing whether a summary preserves the gist of the source text; i.e. the core information relayed by the source text. This is challenging, as paraphrased texts obfuscate the connection between the information contained across the source and summary texts. Students can commit mistakes by misunderstanding the original text, by outright fabricating new information or by misreporting facts (Winograd, 1984). Teachers have to read a summary multiple times to correctly assess the level of content overlap between two texts. In a classroom setting, each teacher adopts their own personal technique for summary evaluation. Extensive research has been conducted in the applied linguistics (AL) field to develop a framework for the evaluation and analysis of summary protocols; i.e. student responses to a summarization task (Johns and Mayes, 1990; Yu, 2007; Li, 2014). This framework allows researchers to measure the content overlap between the two documents and assess the comprehension skills of students. The procedure consists of a two-step process: (1) Summaries and their corresponding source texts are divided into short chunks of text conveying information. (2) These chunks are referenced across the texts and paired according to their meaning. Depending on the requirements of the experiment, researchers can later analyse and label the relationship between paired chunks. We refer to the first step as **segmentation** and the second as **alignment**. Although popular, this style of summary response analysis requires particular effort and is a time-consuming task in itself (Bernhardt, 1991; Sawaki, 2005), barring its use from classroom settings. However, such a strict and well-defined procedure is a prime candidate for computerisation.

The objective of this paper is to automatise the summary protocol analysis procedure as detailed in the AL field. The goal is to simplify the researchers' work by providing new tools for automatise the above two processes: segmentation and alignment. These tools can also benefit teachers in evaluating students' reading comprehension as represented by the content of summaries they produce.

Our contribution is the following: (1) We compared different segmentation units from the AL field. (2) We selected the Idea Unit (IU) as the best candidate for this study. (3) We revised the definition of IU and propose a new annotation guideline. (4) We manually annotated a set of student summaries into IUs. (5) We developed IUExtract, an automatic IU segmentation algorithm based on the revised annotation guideline and

tested it on the dataset. (6) We tested an existing alignment algorithm with state of the art models. (7) We propose an online tool for collecting alignment data called SAT.

As language resources, we released to the public a gold standard Idea Unit corpus, the automatic segmentation algorithm and the online alignment annotation tool[1].

2. Related Work

Segmentation is a well-known task in the natural language processing (NLP) field, and several segmentation tasks are discussed in the literature.

Topic segmentation aims to extract text spans conveying information about a shared subject called "topic". Neural network approaches to topic segmentation have been proposed by Glavas et al. (2016), achieving a P_k (Beeferman et al., 1999) between 5.6 and 9.6 on the Choi (2000) dataset and 28.08 on the manifesto dataset[2]. Another approach by Koshorek et al. (2018) measures P_k in the range between 18.3 and 41.63 on a variety of datasets, including Choi (2000). Vinotheni.C and LakshmanaPandian.S (2021) proposed a model combining a Fast Recurring Neural Network with a Bi-LSTM to achieve a maximum of 93.7 in precision and 91.3 in F_1 score.

Another segmentation task is the identification of Elementary Discourse Units (EDUs) in a text. The EDU was introduced by Carlson et al. (2001a) along with the RST Discourse Treebank (RST-DT) corpus (Carlson et al., 2002) to provide resources and incite new research on Rhetorical Structure Theory (RST) (Mann and Thompson, 1988) by reducing inconsistencies in definition across studies. In general, an EDU is a clause, but a list of exceptions and special cases is given. The full annotation guideline was published by Carlson et al. (2001b). Some examples of automatic EDU segmentation algorithms are the statistical-based SPADE (Soricut and Marcu, 2003), the rule-based SLSeg (Tofiloski et al., 2009) and the neural network-based SegBot (Li et al., 2018). Saveleva et al. (2021) proposed SegFormers, a transformer based EDU segmentation model that achieved 97.09 in F_1 score over the English RST-DT dataset. Despite their similarities, EDUs are not exactly the same as IUs and the above-mentioned models cannot be used as-is for IU segmentation. Furthermore, a lack of gold standard data prevents us from training machine learning models for IU segmentation.

3. Segmentation

To ensure the reliability of the automatic segmentation algorithm, it is important to fix the segmentation unit of choice and define the annotation guidelines.

Several segmentation units are adopted in the literature. In this section, we will discuss some of them and choose the best one for automation.

[1] https://tt-cl.github.io/iu-resources/

[2] https://manifestoproject.wzb.eu/

Propositions A proposition is a semantic unit that consists of one or more assertions regarding at least one major argument (Sato, 1988). Propositions have been used in comprehension studies targeting both first language (Connor, 1984; Coffman, 1994) and second language (Connor, 1984; Barnett, 1986) learners. Propositions can vary in length, and can even contain other propositions as dependants (Meyer, 1975). From a computational point of view, propositions are too vague and differ greatly in terms of complexity, from simple lexical propositions – sentences containing only a noun and a verb phrase – to complex propositions with subordinate propositions of their own (Meyer, 1975). These issues result in unstable annotation, as each study will segment propositions differently.

T-Units The *minimally terminable unit*, shortened to T-Unit (Hunt, 1965; Hunt, 1966; Hunt, 1970), is a portion of text composed of a main clause and any other subordinate clauses attached to it. T-Units have been adopted to measure second language learner fluency (Cooper, 1976; Ishikawa, 1995), accuracy (Hirano, 1991; Homburg, 1984; Casanave, 1994) and grammatical complexity (Hunt, 1965; Casanave, 1994). T-Units are defined in a simple manner and humans can perform annotation with relative ease. T-units are word sequences that can be transformed into sentences by capitalising the first word and appending a full stop at the end of the sequence. T-Units pose a problem for the analysis of student summaries, i.e. they are too long. T-Units can contain multiple propositions and convey multiple information in a single unit. This poses additional challenges when aligning information across two different texts, as one unit in the summary could contain multiple pieces of information referring to several other units in the source text.

Idea Units The Idea Unit (IU) is a "chunk of information which is viewed by the speaker/writer cohesively as it is given surface form" (Kroll, 1977). It is described as a semantic chunk but its segmentation rule-set is defined syntactically (Kroll, 1977). The IU has been used for both listening comprehension (Winke et al., 2013; Shin et al., 2016) and written recall studies (Lee and Riley, 1990). IUs have also been adopted in studies analysing non-English texts (Lee, 1986; Lee and Riley, 1990; Ableeva and Lantolf, 2011). Multiple revisions of the annotation guidelines have been proposed throughout the years (Kroll, 1977; Johns, 1985; Johns and Mayes, 1990; Carrell, 1985; Carrell, 1992; Ikeno, 1996; Gecchele et al., 2019). The slight differences between them result in datasets that are not compatible across studies using different guidelines. Furthermore, no reference IU dataset is available to the public. Despite these issues, the syntactic nature of the IU rule-set allows for automation via standard syntactic parsers.

In this paper, we follow the summary analysis procedure detailed by Johns and Mayes (1990) and adopt the Idea Unit (IU) as the segmentation unit. We will first

refine the annotation guideline to be as generic as possible. We will then describe a new IU gold standard corpus that we collected and released to the research community along with this paper. Lastly, we will detail the automatic segmentation algorithm developed on our revision of the IU annotation guidelines.

4. Revision of the Idea Unit Annotation Guidelines

Before Idea Unit segmentation can be computerised, a standardised rule-set needs to be agreed upon.

4.1. Definitions of Idea Unit in the literature

The first definition of IU describes it as a "chunk of information which is viewed by the speaker/writer cohesively as it is given surface form" (Kroll, 1977). This semantic definition is accompanied by a list of seven syntactic rules that compose the annotation guideline. An initial revision by Johns (1985) removes the compound verb rule, reducing the total number of rules to six. A later revision specifies that compound verbs should be separated in different IUs (Johns and Mayes, 1990), unlike in Kroll's guideline, where compound verbs are kept in the same unit.

Carrell provides a different definition, opting to describe the IU in a short paragraph, rather than detailing a list of segmentation rules (Carrell, 1985; Carrell, 1992). The main difference between Kroll and Carrell's IU is that the latter designates "heavy" prepositional phrases as their own IUs. This is sensible, as prepositional phrases can hold information of their own and extracting long prepositional phrases produces shorter IUs closer to atomic semantic units. Ikeno (1996) expanded this definition by specifying what counts as a "heavy" prepositional phrase. Prepositional phrases longer than four words are considered heavy and extracted as an individual Idea Unit. These discrepancies in definition must be overcome for a successful automatic IU segmentation.

4.2. Revision Procedure

In our previous study (Gecchele et al., 2019) we proposed the first revision of the IU annotation guideline. This is an iteration of Kroll (1977) developed with the intent of standardising the definition of IU for automation purposes. However, this version is not comprehensive as it does not mention the long prepositional phrases detailed by Carrell (1992) and Ikeno (1996). Furthermore, this first version is not strict enough, and low inter-annotator agreement can be observed in the annotation, especially when comparing the work of annotators with different backgrounds. This paper further refines the annotation guidelines to improve inter-annotator agreement. We set the following two goals for the revision process.

1. The rule-set must be understandable by humans and allow easily reproducible annotation
2. The rules should be simple enough so that a rule-based algorithm can be developed

We divided ourselves into two teams: two authors with a background in NLP formed the computational team, and two authors with an AL background formed the linguistics team. All authors are non-native English speakers, but they are all professionals in their respective fields.

4.2.1. Dataset

For the revision of the annotation guidelines, a new student summary dataset was collected. The source text, called *Cycloclean*, is an expository text with problem-solution structures embedded in it. The *L2WS 2020 (L2 Written Summary)* corpus is a set of summaries referring to the *Cycloclean* source text collected in 2020. Table 2 shows statistics for the *L2WS 2020* corpus. The summaries were written by 90 students of a linguistics course held at a university in Tokyo, Japan. All the students were non-native English speakers, 88 of which were Japanese first-language speakers. The summaries were collected as part of a course assignment in which the students were asked to read the source text (391 words) and summarise its main ideas and key details in approximately 80 words.

We randomly selected five summaries to form a dev-set. The dev-set was used for the manual revision of the annotation guidelines and as a reference for the development of the rule-based segmentation algorithm detailed in 6. The remaining 85 summaries form the test-set. The test-set was annotated only once with the final revision of IU annotation guidelines. The linguistics team conducted the annotation and produced a double coded dataset by gathering a consensus during a joint meeting. This dataset was used exclusively to test the automatic segmentation algorithm described in 6.

4.2.2. Method

The revision procedure followed a step-wise refinement approach. Both teams annotated a set of five student summaries and their source text. We based the initial annotation on the IU annotation guideline detailed in our previous work (Gecchele et al., 2019). This initial revision is reported in Appendix A.

Each team compared annotations within the team to produce an agreed team annotation. Then, they are compared and analysed by all annotators to revise the guidelines. The revised guidelines were used for the next iteration.

Rule	"κ"	WDiff 3	WDiff 5	P_k 3	P_k 5
Original	0.547	0.199	0.289	0.193	0.254
Revision	0.785	0.142	0.219	0.135	0.168

Table 1: Inter-annotator agreement over the revision of the annotation guidelines. "Original" refers to the annotation guideline detailed in Gecchele et al. (2019).

At each iteration of the revision process, the inter-annotator agreement was measured. This was done by coding each document as a binary string where spaces between tokens coinciding with segment boundaries are represented by a 1, 0 otherwise. Our IAA metrics are Cohen's κ (Cohen, 1960), P_k (Beeferman et al., 1999) and WindowDiff (Pevzner and Hearst, 2002).

The revision process was repeated until an inter-annotator agreement of 0.785 of Cohen's κ was achieved (Table 1).

Once the rule-set revision was agreed upon, the linguistics team proceeded to annotate the remaining 85 documents in the test-set. These were used exclusively to test the automatic segmentation algorithm detailed in 6.

4.3. Proposed Idea Unit Annotation Guidelines

The resulting annotation guideline is composed of ten rules, maintaining the rule approach proposed initially by Kroll (1977). An example of Idea Unit segmentation is shown in Figure 1.

Sentence	Polly, Grace's dog, was professionally trained
IU1	Polly, [...] was professionally trained
IU2	Grace's dog,

Figure 1: An example of a sentence segmented in two Idea Units. IU1 is a discontinuous Idea Unit.

Some of the main changes from previous versions are:

1. Several rules have been rewritten and reordered. A detailed description of appositives and semantically independent phrases is given. Finally, some word-level details (Rule 10 in Figure 2) are given to increase agreement.
2. Following Carrell (1992) and Ikeno (1996), long and semantically independent prepositional phrases count as an Idea Unit (Rule 8 in Figure 2). This was done to keep the IUs as short as possible.
3. Diverging from Johns and Mayes (1990), we follow the original definition of IU (Kroll, 1977) and we opt to maintain compound verbs in the same idea unit (Rule 4 in Figure 2). This was done because separating compound verbs led to a loss of information during annotation, as certain facts could only be inferred by stitching together the IUs containing the compound verbs.
4. Following our previous work (Gecchele et al., 2019), we specified that IUs can be discontinuous or continuous. We say an Idea Unit is discontinuous when it is composed of words that are not adjacent in the text. See Figure 1.
5. To improve upon our previous work Gecchele et al. (2019), we provide examples of phrases that are considered set off from their sentences (Rules 3.3 and 3.5 in Figure 2).

1. A subject and verb count as one idea unit together with (when present) a
   (a) direct object,
   (b) short prepositional phrase,
   (c) adverbial element,
   (d) mark of subordination,
   (e) a combination of the above.
2. Subordinate clauses, full relative clauses and reduced relative clauses count as separate idea units.
3. Phrases that are set off from the sentence with commas are counted as separate idea units. We define a phrase to be "set off" from its sentence when they interrupt or shift the focus of the discourse.
   3.1. Parenthetical expressions – phrases set off with parentheses, hyphens or other punctuation marks - should also be counted as separate idea units.
   3.2. Appositives by definition are set off from the discourse and should be split into separate Idea Units.
   3.3. Adverbial conjunctions that do not add meaningful information (e.g.: "However,") are not to be split into separate Idea Units.
   3.4. Citations are counted as separated idea units only when they are set off from the sentence in their entirety.
   3.5. Temporal adverbial modifiers and prepositional phrases that relay temporal information are split into separate Idea Units when they are located at the beginning of a sentence, even if they are not followed by a punctuation mark (e.g.: "In 2015,").
4. Verbs whose structure requires or allows a multiple auxiliaries are counted with all their verbal elements as one idea unit.
5. Infinitive clauses that modify a noun or adverb count as one idea unit.
6. Other types of elements that count as idea units are
   6.1. Absolutes and
   6.2. Verbals that define purpose or scope – infinitives that can be prefixed by "in order to"
7. Idea Units can be discontinuous – an idea unit can be composed of segments of texts that are not directly adjacent to each other.
8. Semantically independent prepositional phrases that are long in length are counted as one Idea Unit. The limit between long and short prepositional phrases is left to the judgement of the researcher adopting the rule-set.
9. Each rule is equally important. Idea Units should always be segmented to be the smallest size as possible, regardless of rule order.
10. Word level details:
    10.1. Subordinating conjunctions and relative pronouns are always attached to the subordinate clause.
    10.2. Punctuation is always attached to the word to the left, with the exception of open parentheses which are attached to the right.

Figure 2: Revised Idea Unit annotation guidelines.

5. Idea Unit Gold Standard Corpus

As part of this paper, we release an Idea Unit gold standard annotation corpus as a new language resource[3]. The *L2WS 2021* corpus is a set of summaries referring to the *Cycloclean* source text described in 4.2.1. The

[3]https://tt-cl.github.io/iu-resources/

```
Rules ← [Booleanf]                          ▷ A list containing each annotation guideline as a boolean function
function IUEXTRACT(Sentence)                ▷ The wrapper function
    Tree ← Spacy(Sentence)                  ▷ Generate dependency Tree
    Tagged ← Queue()                        ▷ Initialize the empty queue
    TAG(Tree.head(), Tagged)                ▷ Populate the queue with the head nodes of IUs
    INDEX(Tagged)                           ▷ Assign IU indexes
    return Tree
function TAG(Node, Queue)                   ▷ Tag each node with the appropriate rule
    for each Child ∈ Node.children() do
        TAG(Child, Queue)                   ▷ Do the children first, DFS
    for each Rule ∈ Rules do
        if Rule(Node) = True then           ▷ One rule is applicable
            Queue.push(Node)                ▷ Tag the node for segmentation
            Node.tags ← Rule.name()         ▷ Append the rule name to the node
function INDEX(Queue)                       ▷ Assign an IU index to each word
    Queue.reverse()                         ▷ The tagged queue is reversed to traverse the tree bottom up
    CurrentIndex ← 0                        ▷ Initialize the first index
    for each TaggedNode ∈ Queue do
        CurrentIndex ← CurrentIndex + 1     ▷ Each IU has a different index
        TaggedNode.IUindex ← CurrentIndex
        for each Child ∈ TaggedNode.children() do
            if Child.IUindex = NULL then    ▷ The node was not yet indexed
                Child.IUindex ← CurrentIndex
```

Figure 3: Pseudocode of IUExtract: the Idea Unit Segmentation Algorithm

L2WS

	Pub.	#Docs	#Tokens	#IUs
Source	YES	1	391	49

Set	Pub.	#Sum.	#Avg Tokens	#Avg IUs
2020 dev	NO	5	109.0	16
2020 test	NO	85	98.9	13.8
2021	YES	40	94.4	12.8

Table 2: Statistics for the **L2WS** corpora. The "Pub". column refers to whether the corpus will be published. The "Avg Words" column gives the average number of words per summary, while "Avg IUs" gives the average number of IUs per summary.

data was collected in 2021 with the intent of releasing an Idea Unit gold standard to the public. This was necessary, as *L2WS 2020* was collected with consent for research use, but no consent was given for publishing the data. For *L2WS 2021*, we obtained consent from participants for research use and sharing of the collected data. The data collection procedure was also reviewed and approved by the University research ethics committee. The summaries were collected as part of a the same course assignment described for the *L2WS 2020* corpus. *L2WS 2021* is comprised of 40 summaries written by 40 university students. All the students speak Japanese as a first language. The linguistics team annotated the corpus according to the Idea Unit annotation guideline detailed in this paper. The corpus is double coded via consensus between the annotators. Corpus statistics are presented in Table 2.

6. IUExtract: Automatic IU Segmentation Algorithm

Taking a machine learning approach for segmentation is difficult due to the limited amount of available training data. Instead, we opted to translate the annotation guidelines into a rule-based segmentation algorithm. This method has an additional advantage; rule-based algorithms can be directly mapped on the annotation guidelines and, therefore, can be interpreted by humans. This will allow researchers to analyse the segmentation algorithm and easily adapt it to suit any specific requirement. Such alterations are not easy to implement when using black-box machine learning techniques.

This algorithm, which we refer to as "IUExtract", is released to the public as a python package[4]. Figure 3 shows a short pseudo-code of IUExtract.

6.1. Implementation

Each extractive rule from the annotation guideline was coded as a Boolean function that takes as input a word token and returns a positive result if this word and all its dependants should be identified as an IU. To compute these judgements, the texts are first parsed into dependency trees via SpaCy (Honnibal and Johnson, 2015).

IUExtract parses and segments texts sentence by sentence. For each sentence, the dependency tree is explored via a depth-first search. Each node is tested

[4]https://tt-cl.github.io/iu-resources/

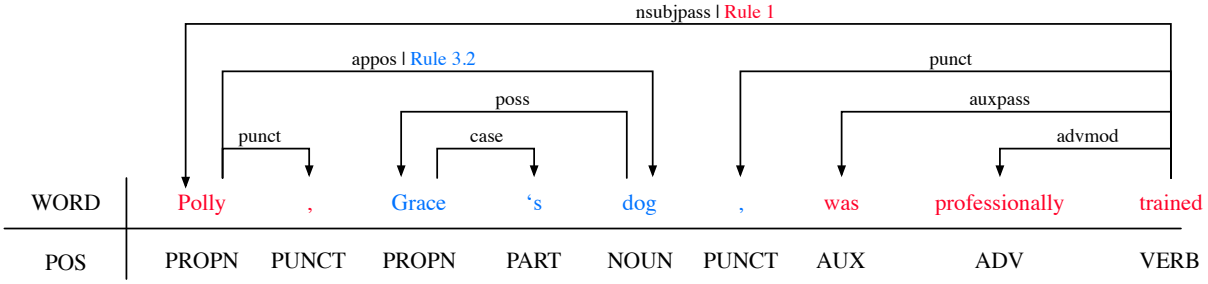

Figure 4: An example of the segmentation algorithm, showing how discontinuous IUs can be discovered by extracting appositions (rule 3.2). First, the passive subject and apposition dependency arcs are tagged for segmentation as they respectively fall under **rule 1** and **rule 3.2**. In the picture, the colours red and blue are assigned to the two IUs to help distinguish them. The head node of each IU is signalled by a coloured dependency label. The noun "*dog*" is closer to the leaves when compared with the verb "trained" and as such it is tagged first along with its dependants. The remaining words are coloured in red, as they all depend on the verb "*trained*". Lastly, the punctuation (commas) is attached to the previous token as per **rule 10.2**.

against the list of segmentation rules. If a Boolean function returns true, the corresponding rule number is stored inside the token. This step is called *tagging*. Each tagged node is put into a processing queue.
After tagging, the algorithm proceeds to *indexing*. The processing queue is reversed to explore the tree again, this time in a bottom-up fashion. Each node from the queue and all of its children are assigned a unique IU index. If a visited node already has an IU index it is left unchanged, as the node was already visited when exploring the children of a previous node in the queue. Once indexing is complete, all the words with the same index can be joined to form an IU. An example of how the algorithm works is shown in Figure 4.

6.2. Evaluation

IUExtract was evaluated in terms of Precision, Recall and F_1 score on the *L2WS 2020* test-set. Numbers for simple agreement, perfect segment match and IU length are also included in the analysis. Later, the algorithm was also tested on the *L2WS 2021* corpus. The formulas for Precision, Recall F_1 score are the following:

$$Precision = \frac{|AutoBoundaries \cap GoldBoundaries|}{|AutoBoundaries|},$$

$$Recall = \frac{|AutoBoundaries \cap GoldBoundaries|}{|GoldBoundaries|},$$

$$F_1 = 2 * \frac{Precision \times Recall}{Precision + Recall},$$

where $AutoBoundaries$ is the set of Idea Unit boundaries automatically extracted by the algorithm and $GoldBoundaries$ is the set of manually annotated segment boundaries.
The boundaries are counted via a binary representation of each document. Precision, Recall and F_1 only measure agreement over segment boundaries. For this reason, we count Perfect Disc. IUs; i.e. discontinuous IU pairs that match exactly on a word level basis regardless of the segment boundaries.

	L2WS			
	2020 Test-set		*2021*	
	IUExtract	Gold	IUExtract	Gold
#IUs	1264	1174	542	512
#Disc. IUs	74	67	33	26
P IUs	723	–	305	–
P Disc. IUs	22	–	8	–
P IU ratio	0.572	–	0.563	–
P Disc. IU ratio	0.297	–	0.242	–
AVG IU length	6.649	7.158	6.967	7.375
IU length VAR	10.59	10.27	12.06	10.73
Precision	0.800	–	0.789	–
Recall	0.868	–	0.844	–
F_1 Score	0.833	–	0.815	–

Table 3: Evaluation results for the segmentation algorithm. The italicised P stands for Perfect. The ratios in the rows "P IU ratio" and "P Disc. IU ratio" are calculated by dividing the number of perfect Idea Units by the number of automatically extracted IUs. Average IU length, variance, Precision, Recall and F_1 score are all micro-averaged.

The evaluation results are reported in Table 3. The figures show promising results, both in terms of precision and recall, reaching a 0.800 in precision and 0.868 in recall over the test-set. Although they are slightly lower, similar figures are measured on *L2WS 2021*, the corpus collected months after the development of the algorithm. In the gold standard, the average IU length increases from the 7.158 measured over *L2WS 2020* to 7.375 over *L2WS 2021*. A comparable increment is observed when looking at IUExtract figures; measuring 6.649 over *L2WS 2020* and 6.967 over *L2WS 2021*. Finally, the segmentation algorithm produces more Idea Units than the gold standard, leading to units of shorter average length. However, the percentage of perfect

matches between automatically extracted Idea Units and the reference gold standard is high, figuring at 56.3%.

6.3. Comparison against the EDU

This paper aims to automate the summary analysis procedure from the AL field via Idea Units. However, the IU is similar to the Elementary Discourse Unit (EDU) (Carlson et al., 2001a) from the NLP field. A comparison between the two is of interest. These two units cannot be compared directly by comparing gold standards, as the annotation guidelines are different. For instance, appositions are not extracted as EDUs, but they are considered individual IUs. However, similarities also exist. EDUs opt to extract long prepositional phrases as individual segments, following closely the IU annotation guidelines detailed in Carrell (1992) and Ikeno (1996) and our new revision.
As we do not have access to gold standard datasets annotated according to both IU and EDU guidelines, we leave an in-depth comparison of the two segmentation techniques as future work.

7. Alignment

In our previous study (Gecchele et al., 2019), we proposed an algorithm for the automatic alignment of IUs. The algorithm relies on Stanford's GloVe pre-trained word embeddings (Pennington et al., 2014) to map words into a vector space according to their meaning. An IU vector is calculated by averaging the word embeddings for each word of an Idea Unit. For each Summary IU, a Prediction Ranking is computed by listing each Source IU in descending order of cosine similarity. In this paper, we updated the GLoVe embedding model with the state-of-the-art SBERT (Reimers and Gurevych, 2019) and tested the alignment algorithm against our previous version. We also propose a tool to simplify the collection of alignment gold standards to develop new models.

7.1. Alternative Word-embedding Models

The following embedding models are explored:

- **GloVe** (Pennington et al., 2014): Stanford's embedding model pre-trained on the Wikipedia 2014 + Gigaword 5 corpus. This is the baseline model as it is the best performing approach from our previous study (Gecchele et al., 2019).
- **Word2Vec**: SpaCy (Honnibal and Johnson, 2015)'s pre-trained model "en_core_web_lg" is a word2vec (Mikolov et al., 2013) model pre-trained on the GloVe Common Crawl and OntoNotes 5.0 (Weischedel et al., 2013) datasets.
- **SBERT**: BERT (Devlin et al., 2019) allows researchers to feed pairs of sentences to the transformer network and retrieve a similarity value that, unlike traditional word embeddings, is context-aware. Sentence BERT (Reimers and Gurevych, 2019) is the state-of-the-art for the generation of sentence embeddings. Of the provided pre-trained models, "paraphrase-mpnet-base-v2" was selected, as it is the best in terms of performance[5].

7.2. Data

	Source 1	Summary	Source 2	Summary
Words	996	185.5	807	204.5
Avg #IUs	111	20.6	89	24.6
Avg #Links	—	18.0	—	21.3

Table 4: Statistics for the alignment data sets. The "Avg #Links" row shows the average number of aligned IUs across the summaries and their source text.

To preserve consistency, the experiment was run over the same dataset used in the preliminary study. The dataset consists of 20 summaries written by ten PhD students at a university in the UK. The students were given two texts from the IELTS test and they were instructed to summarise them to 25% of the length of the original text. The topics of the passages were "the preservation of endangered languages" and "the impact of noise on cognitive abilities".
The dataset statistics are shown in Table 4.The 20 summaries of the combined dataset were manually segmented according to the older annotation guideline (Gecchele et al., 2019) by the computational team. A double coded gold standard was compiled via consensus. This dataset is accompanied by annotation data compiled manually by the same annotators.

7.3. Results

The results are shown in Figure 5. SBERT is the best model, improving both precision and recall. Word2Vec achieves similar performance to GloVe, as the only difference between the models is a larger training corpus for Word2Vec. Given that the objective is to pair Idea Units automatically, an optimal algorithm would rank the IU pairs linked in the gold standard at the highest spot on the Prediction Ranking. This is reflected by the numbers obtained with a window size of one. At threshold 1, the GloVe baseline achieves 0.332 in precision and 0.366 in recall. SBERT achieved slightly better, showing 0.375 in precision and 0.415 in recall. Even the state-of-the-art SBERT model provides insufficient performance for a reliable alignment, suggesting that further improvement is required.

8. Alignment Data Collection Tool

Alignment gold standard data is difficult to collect. While annotators can intuitively produce segmentation

[5] www.sbert.net/docs/pretrained_models.html

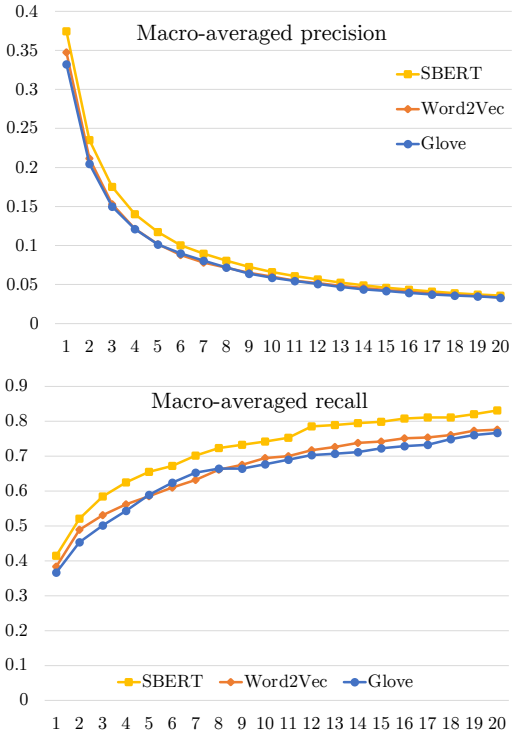

Figure 5: Evaluation results for the alternative word embedding models. The x axis indicates the size of the window of predictions; the number of top n IU pairs retrieved from the "Prediction Ranking".

gold standards by breaking sentences into multiple segments, they have to link IU pairs across documents when they are working with alignment. Annotators have to find combinations of segments across texts and annotate the indexes of matching IUs in a list of tuples. Since IUs can be aligned on a many-to-many basis, the manual annotation procedure is challenging and time-consuming.

To reduce the chance of mistakes and facilitate the collection of alignment data, we developed an online tool called SAT – Segmentation and Alignment Tool. The source code is available online[6] and can be easily deployed on a private server to conduct alignment annotation. The website shows summaries and their source text side by side, both automatically segmented in IUs via IUExtract. The IUs are shown as bubbles; annotators can manually link IUs by clicking on the bubbles through the GUI. A screenshot of the alignment GUI is shown in Appendix B. SAT is an extension of Segment Matcher, our previous tool described in (Gecchele et al., 2019), as SAT can automatically segment raw texts into IUs and does not require annotators to conduct manual segmentation in advance. Annotators can modify the IU boundaries manually to correct segmentation errors through the GUI. The error corrections are recorded in a log file for further refinement of the segmentation algorithm. The log file and the annotation are periodically sent to the back-end and stored on the database. This allows researchers to automatically collect new alignment data by hiring external annotators and directing them to the website.

[6]https://tt-cl.github.io/iu-resources/

9. Conclusion and Future Work

In this paper, we developed tools that can be used by applied linguistics researchers to assess students' comprehension skill through the analysis of summary composition. In this field, researchers have analysed whether a summary preserves the gist of the corresponding source text by matching semantic chunks of information (i.e. Idea Units) across a summary and its source text.

We revised the existing guidelines for Idea Unit annotation to improve the inter-annotator agreement. The revision improved the IAA from 0.547 to 0.785 of Cohen's κ.

We collected L2WS 2021, a novel corpus comprised of 40 summaries composed by second language learners at a university. Each summary was annotated according to the revised guidelines. This corpus is released to the public as a novel language resource.

Next, we developed an automatic segmentation algorithm, IUExtract, following the revised IU rule-set. This algorithm was constructed by implementing each rule in the guidelines as a boolean function that relies on a dependency parser to compute its judgements. Our results show that the segmentation algorithm retains its performance across datasets collected months apart. The F_1 score of 0.833 recorded on the *L2WS 2020* is slightly higher than the 0.815 recorded across the *L2WS 2021* corpus.

We tested newer embedding models on an existing IU alignment algorithm. Our results showed only a slight increase in precision for the top 1 alignment pairs, raising to a 0.375 with SBERT (Reimers and Gurevych, 2019). We deemed this figure insufficient for an effective alignment. For this reason, we developed SAT, an online Segmentation and Alignment Tool that can be used to collect new alignment gold standards easily.

In the future, we plan to use the annotation tool to gather extensive alignment gold standard data and develop machine learning solutions for automatic alignment. We also plan on comparing the Idea Unit against the Elementary Discourse Unit both empirically and theoretically. First, we will compare the two annotation guidelines rule by rule. Next, we will manually annotate part of the RST-DT dataset (Carlson et al., 2002) into IUs and the *L2WS 2021* corpus into EDUs, allowing for a direct comparison of the two guidelines.

Acknowledgement

This work was supported by the Japan Society for the Promotion of Science (JSPS) Grant-in-Aid for Scientific Research (B) (No. 20H01292; PI: Yasuyo Sawaki) This work was supported by JST SPRING, Grant Number JPMJSP2106.

10. Bibliographical References

Ableeva, R. and Lantolf, J. (2011). Mediated dialogue and the microgenesis of second language listening comprehension. *Assessment in Education: Principles, Policy and Practice*, 18(2):133–149.

Barnett, M. A. (1986). Syntactic and Lexical/Semantic Skill in Foreign Language Reading: Importance and Interaction. *The Modern Language Journal*, 70(4):343–349, dec.

Beeferman, D., Berger, A., and Lafferty, J. (1999). Statistical models for text segmentation. In *Machine Learning*, volume 34, pages 177–210.

Bernhardt, E. B. (1991). *Reading development in a second language: theoretical, empirical, and classroom perspectives*. Ablex Publishing Corporation, Norwood, New Jersey.

Carlson, L., Marcu, D., and Okurowski, M. E. (2001a). Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory. In *Proceedings of the Second SIGdial Workshop on Discourse and Dialogue*, volume 16, pages 1–10, Morristown, NJ, USA. Association for Computational Linguistics.

Carlson, L., Marcu, D., and Okurowski, M. E. (2001b). Building a discourse-tagged corpus in the framework of rhetorical structure theory. In *Proceedings of the Second SIGdial Workshop on Discourse and Dialogue - Volume 16*, SIGDIAL '01, page 1–10, USA. Association for Computational Linguistics.

Carrell, P. L. (1985). Facilitating ESL Reading by Teaching Text Structure. *TESOL Quarterly*, 19(4):727, dec.

Carrell, P. L. (1992). Awareness of Text Structure: Effects on Recall. *Language Learning*, 42(1):1–18.

Casanave, C. P. (1994). Language development in students' journals. *Journal of Second Language Writing*, 3(3):179–201.

Choi, F. Y. Y. (2000). Advances in domain independent linear text segmentation. In *Proceedings of the 1st North American Chapter of the Association for Computational Linguistics Conference*, NAACL 2000, page 26–33, USA. Association for Computational Linguistics.

Coffman, G. A. (1994). The influence of question and story variations on sixth graders' summarization behaviors. *Reading Research and Instruction*, 34(1):19–38, sep.

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37–46.

Connor, U. (1984). Recall of Text: Differences between First and Second Language Readers. *TESOL Quarterly*, 18(2):239, jun.

Cooper, T. C. (1976). Measuring Written Syntactic Patterns of Second Language Learners of German. *The Journal of Educational Research*, 69(5):176–183, jan.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, jun. Association for Computational Linguistics.

Gecchele, M., Yamada, H., Tokunaga, T., and Sawaki, Y. (2019). Supporting content evaluation of student summaries by idea unit embedding. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 343–348, Florence, Italy, August. Association for Computational Linguistics.

Glavas, G., Nanni, F., and Ponzetto, S. P. (2016). Unsupervised text segmentation using semantic relatedness graphs. **SEM 2016 - 5th Joint Conference on Lexical and Computational Semantics, Proceedings*, pages 125–130.

Graham, S. and Perin, D. (2007). A Meta-Analysis of Writing Instruction for Adolescent Students. *Journal of Educational Psychology*, 99(3):445–476.

Hirano, K. (1991). The Effect of Audience on the Efficacy of Objective Measures of EFL Proficiency in Japanese University Students(Second Language Acquisition Theory & Analysis of Scholarstic Abilities). *ARELE: Annual Review of English Language Education in Japan*, 2(2):21–30.

Homburg, T. J. (1984). Holistic Evaluation of ESL Compositions: Can It Be Validated Objectively? *TESOL Quarterly*, 18(1):87, mar.

Honnibal, M. and Johnson, M. (2015). An Improved Non-monotonic Transition System for Dependency Parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1373–1378, Lisbon, Portugal, sep. Association for Computational Linguistics.

Hunt, K. W. (1965). Grammatical structures written at three grade levels. In *National Council of Teachers of English, Research Report No. 3*, pages 1–176.

Hunt, K. W. (1966). Recent Measures in Syntactic Development. *Elementary English*, 43(7):732–739.

Hunt, K. W. (1970). Syntactic Maturity in Schoolchildren and Adults. *Monographs of the Society for Research in Child Development*, 35(1):iii—-67, feb.

Ikeno, O. (1996). The effects of text-structure-guiding questions on comprehension of texts with varying linguistic difficulties. 大学英語教育学会紀要, 1975:51–68.

Ishikawa, S. (1995). Objective measurement of low-proficiency EFL narrative writing. *Journal of Second Language Writing*, 4(1):51–69, jan.

Johns, A. M. and Mayes, P. (1990). An analysis of summary protocols of university ESL students. *Applied Linguistics*, 11(3):253–271, sep.

Johns, A. M. (1985). Summary Protocols of "Underprepared" and "Adept" University Students: Repli-

cations and Distortions of the Original. *Language Learning*, 35(4):495–512.

Keck, C. (2014). Copying, paraphrasing, and academic writing development: A re-examination of L1 and L2 summarization practices. *Journal of Second Language Writing*, 25(1):4–22, sep.

Koshorek, O., Cohen, A., Mor, N., Rotman, M., and Berant, J. (2018). Text Segmentation as a Supervised Learning Task. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 469–473, New Orleans, Louisiana, jun. Association for Computational Linguistics.

Kroll, B. (1977). Combining ideas in written and spoken English: a look at subordination and coordination. In Elinor Ochs et al., editors, *Discourse across time and space*, volume 5 of *S.C.O.P.I.L.* Los Angeles, Calif.: Dept. of Linguistics, University of Southern California.

Lee, J. F. and Riley, G. L. (1990). The effect of prereading, rhetorically-oriented frameworks on the recall of two structurally different expository texts. *Studies in Second Language Acquisition*, 12(1):25–41.

Lee, J. F. (1986). Background Knowledge & L2 Reading. *The Modern Language Journal*, 70(4):350–354.

Li, J., Sun, A., and Joty, S. (2018). SegBot: A Generic Neural Text Segmentation Model with Pointer Network. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, pages 4166–4172, California, jul. International Joint Conferences on Artificial Intelligence Organization.

Li, J. (2014). Examining genre effects on test takers' summary writing performance. *Assessing Writing*, 22:75–90, oct.

Mann, W. C. and Thompson, S. A. (1988). Rhetorical Structure Theory: Toward a functional theory of text organization. *Text - Interdisciplinary Journal for the Study of Discourse*, 8(3).

Meyer, B. J. F. (1975). Identification of the Structure of prose and Its Implications for the Study of Reading and Memory. *Journal of Reading Behavior*, 7(1):7–47, mar.

Mikolov, T., Chen, K., Corrado, G., and rey Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. In C J C Burges, et al., editors, *NIPS'13: Proceedings of the 26th International Conference on Neural Information Processing Systems*, volume 2, pages 3111–3119, Red Hook, NY, USA. Curran Associates, Inc.

Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Pevzner, L. and Hearst, M. A. (2002). A critique and improvement of an evaluation metric for text segmentation. *Computational Linguistics*, 28(1):19–36, mar.

Reimers, N. and Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China, nov. Association for Computational Linguistics.

Sato, C. J. (1988). Origins of complex syntax in interlanguage development. *Studies in Second Language Acquisition*, 10(3):371–395.

Saveleva, E., Petukhova, V., Mosbach, M., and Klakow, D. (2021). Discourse-based Argument Segmentation and Annotation. In *Proceedings of the 17th Joint ACL - ISO Workshop on Interoperable Semantic Annotation*, pages 41–53, Groningen, The Netherlands (online), jun. Association for Computational Linguistics.

Sawaki, Y. (2005). The generalizability of summarization and free recall ratings in L2 reading assessment. *JLTA Journal*, 7:21–44.

Shin, S.-Y., Lidster, R., Sabraw, S., and Yeager, R. (2016). The effects of L2 proficiency differences in pairs on idea units in a collaborative text reconstruction task. *Language Teaching Research*, 20(3):366–386, may.

Soricut, R. and Marcu, D. (2003). Sentence level discourse parsing using syntactic and lexical information. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 149–156.

Tofiloski, M., Brooke, J., and Taboada, M. (2009). A syntactic and lexical-based discourse segmenter. In *ACL-IJCNLP 2009 - Joint Conf. of the 47th Annual Meeting of the Association for Computational Linguistics and 4th Int. Joint Conf. on Natural Language Processing of the AFNLP, Proceedings of the Conf.*, ACLShort '09, pages 77–80, USA. Association for Computational Linguistics.

Vinotheni.C and LakshmanaPandian.S. (2021). Deep Learning-Based Text Segmentation in NLP Using Fast Recurrent Neural Network with Bi-LSTM. In *Advances in Parallel Computing*, volume 38, pages 87–93. Elsevier, oct.

Westby, C., Culatta, B., Lawrence, B., and Hall-Kenyon, K. (2010). Summarizing Expository Texts. *Topics in Language Disorders*, 30(4):275–287, oct.

Winke, P., Gass, S., and Sydorenko, T. (2013). Factors influencing the use of captions by foreign language learners: An eye-tracking study. *Modern Language Journal*, 97(1):254–275, mar.

Winograd, P. N. (1984). Strategic Difficulties in Summarizing Texts. *Reading Research Quarterly*, 19(4):404.

Yu, G. (2007). Students' voices in the evaluation of

their written summaries: Empowerment and democracy for test takers? *Language Testing*, 24(4):539–572.

11. Language Resource References

Carlson, Lynn and Marcu, Daniel and Okurowski, Mary Ellen. (2002). *RST Discourse Treebank, LDC2002T07*. Linguistic Data Consortium, Philadelphia, PA.

Weischedel, Ralph and Palmer, Martha and Marcus, Mitchell and Hovy, Eduard and Pradhan, Sameer and Ramshaw, Lance and Xue, Nianwen and Taylor, Ann and Kaufman, Jeff and Franchini, Michelle and Others. (2013). *Ontonotes release 5.0, LDC2013T19*. Linguistic Data Consortium, Philadelphia, PA.

Appendix

A. Previous Idea Unit Annotation Guideline Revision

In our previous, we proposed an initial revision of the Idea Unit annotation guidelines. This was based off of the guideline proposed in (Kroll, 1977) and no additional sources. In this appendix, we report our first revision detailed in (Gecchele et al., 2019).

1. a subject and verb counted as one idea unit together with (when present) a (a) direct object, (b) prepositional phrase, (c) adverbial element, (d) mark of subordination, or (e) a combination of the above
2. full relative clauses counted as one idea unit when the relative pronoun was present
   (a) phrases that are set off by a complementizer are counted as an Idea Unit
   (b) subordinate conjunctions and relative pronouns are always attached to the subordinate clause
3. phrases which occurred in sentence initial position followed by a comma or which were set off from the sentence with commas were counted as separate idea units
   (a) adverbial conjunctions (e.g.: “However,”) are not to be split into separate Idea Units
   (b) citations are counted as separeted idea units only when they are set off from the sentence in their entirety
4. verbs whose structure requires or allows a verbal element as object were counted with both verbal elements as one idea unit
5. reduced clauses in which a subordinator was followed by a non-finite verb element were counted as one idea unit
6. post-nominal -ing phrases used as modifiers counted as one idea unit
7. other types of elements counted as idea units were (a) absolutes, (b) appositives, and (c) verbals
8. An idea unit can be discontinuous

B. Alignment area of SAT

A screenshot of the alignment area of our Segmentation and Alignment Tool - SAT.

The summary and the corresponding source text are shown side-by-side. Each text is automatically segmented in Idea Units via IUExtract. The IUs are displayed as clickable bubbles. Users are able to align segments by first clicking on a Summary IU and then selecting the corresponding IU in the source text. In the screenshot, the green bubbles represent Summary IUs that have already been manually aligned.