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
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

tt-cl.github.io/iu-resources/

tt-cl.github.io/iu-resources/

OUR CONTRIBUTION

- A **new revision** of the Idea Unit annotation guideline
- An Idea Unit **gold standard dataset**
- An **automatic segmentation algorithm**
- An **online tool** to facilitate alignment data collection

THE IDEA UNIT

In Applied Linguistics, 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). The IU can be used to assess students' listening comprehension and written recall via **segmentation** and **alignment** (Figure 1).

We expand upon our previous work (Gecchele et al., 2019) and release an updated **Idea Unit Annotation Guideline** (Figure 2).

Our tests show that the new annotation guidelines improve the inter-annotator agreement from 0.547 to 0.785 of Cohen's k (Cohen, 1960).

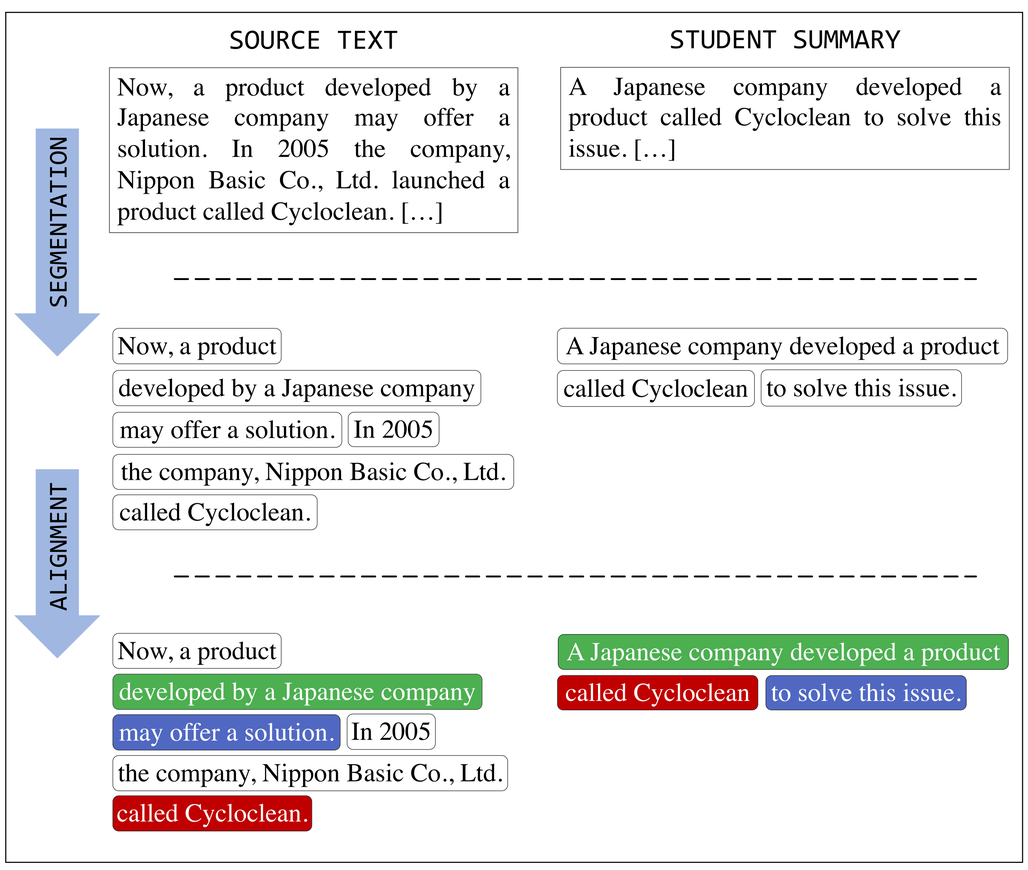

Figure 1: An example of Idea Unit segmentation and alignment.

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: The revision of Idea Unit annotation guidelines.

CORPUS: L2WS 2021

We release an Idea Unit gold standard corpus **L2WS 2021** (L2 Written Summary). The corpus is comprised of 40 summaries written by 40 university students as part of a course assignment. All the summaries refer to a source text that describes a new device that can purify water without electricity. This source text is included in the corpus.

The students were asked to read the source text (391 words) and summarise its main ideas and key details in approximately 80 words. All the students speak Japanese as a first language.

The data is manually annotated according to the IU annotation guidelines released with this paper.

An additional dataset comprised of 80 summaries, **L2WS 2020**, was also collected. However, this dataset cannot be shared with the public due to a lack of consent for sharing from the part of the students. **L2WS 2020** was used exclusively for developing and testing the automatic segmentation algorithm **IUExtract**.

L2WS 2021

	#Docs	#Avg Tokens	# Avg IUs
Source text	1	391	49
Summaries	40	94.4	12.8

Table 1: Statistics for the **L2WS 2021** dataset.

AUTOMATIC SEGMENTATION ALGORITHM: IUEXTRACT

IUExtract is an automatic rule-based segmentation algorithm released as a python package. We developed the algorithm by translating the annotation guidelines into a rule-based segmentation algorithm.

We tested this algorithm against the **L2WS 2020 test set** and **L2WS 2021** corpus. The algorithm was evaluated in terms of Precision, Recall, F_1 score and *Perfect* IU ratio. The formulas for Precision, Recall and 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.

	L2WS			
	2020 Test-set		*2021*	
	IUExtract	Gold	IUExtract	Gold
#IUs	1264	1174	542	512
#Disc. IUs	74	67	33	26
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 2: Evaluation results for the segmentation algorithm. Average IU length, variance, Precision, Recall and F_1 score are all micro-averaged.

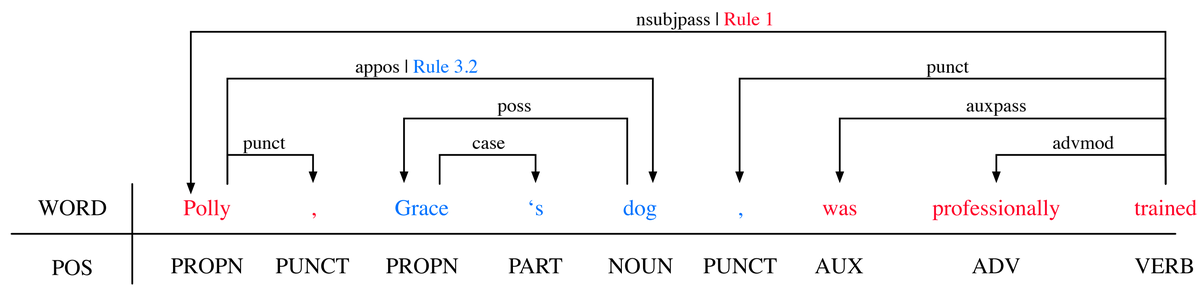

Figure 3: An example of **IUExtract's** functionality. The dependency tree is explored and the arcs "nsubjpass" and "appos" are labelled for segmentation. The satellite of these arcs and each of their children are segmented into an IU.

ALIGNMENT COLLECTION PLATFORM: SAT

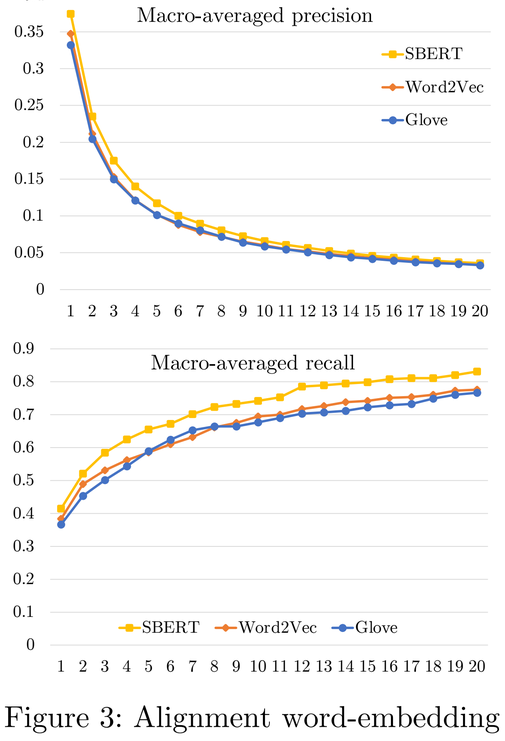

Figure 3: Alignment word-embedding model comparison. The x axis indicates the size of the window of predictions

We tested the alignment algorithm proposed in our previous work (Gecchele et al., 2019) with the more recent word-embedding models. We tested GloVe embeddings (Pennington et al., 2014), SpaCy's Word2Vec implementation (Honnibal and Johnson, 2015) and Sentence BERT (Reimers Gurevych, 2019).

The results are insufficient for effective alignment, with the best model, SBERT, sporting only 0.375 in maximum precision and 0.415 in maximum recall.

We developed a **Segmentation and Alignment Tool – SAT** to facilitate the collection of new alignment gold standard data. **SAT** is a website that can be used by annotators to link Idea Units across texts in a graphical manner.

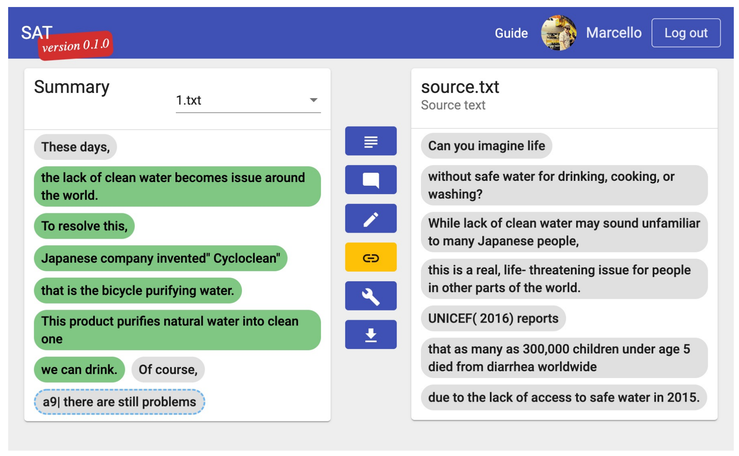

Figure 4: A screenshot of the alignment section of SAT.

AKNOWLEDGEMENTS

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

REFERENCES

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

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.

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.

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.

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.

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.