Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 6884–6888
Marseille, 20-25 June 2022

JADE: Corpus for Japanese Definition Modelling

Han Huang[†], Tomoyuki Kajiwara[‡], Yuki Arase[⋆]
[†]amanotaiga3665@gmail.com
[‡]Graduate School of Science and Engineering, Ehime University
kajiwara@cs.ehime-u.ac.jp
[⋆]Graduate School of Information Science and Technology, Osaka University
arase@ist.osaka-u.ac.jp

Abstract

This study investigated and released the JADE, a corpus for Japanese definition modelling, which is a technique that automatically generates definitions of a given target word and phrase. It is a crucial technique for practical applications that assist language learning and education, as well as for those supporting reading documents in unfamiliar domains. Although corpora for development of definition modelling techniques have been actively created, their languages are mostly limited to English. In this study, a corpus for Japanese, named JADE, was created following the previous study that mines an online encyclopedia. The JADE provides about 630k sets of targets, their definitions, and usage examples as contexts for about 41k unique targets, which is sufficiently large to train neural models. The targets are both words and phrases, and the coverage of domains and topics is diverse. The performance of a pre-trained sequence-to-sequence model and the state-of-the-art definition modelling method was also benchmarked on JADE for future development of the technique in Japanese. The JADE corpus has been released and available online.

Keywords: definition generation, language learning and education, computer-assisted language learning (CALL)

1. Introduction

Definition modelling aims to generate a definition of a word and phrase in context. It is a crucial technology for language learning/education support and reading assistance, such that users can refer to meaning of unfamiliar expressions on the fly. For developing sophisticated definition modelling techniques, a corpus that provides definitions of words and phrases is crucial. Since the meanings of words and phrases are context dependent, such corpus should provide context-aware definitions, *i.e.* a set of word/phrase, its definition, and context in which the word/phrase is used. Additionally, to help language learning and reading, the corpus should be updated with neologisms and trendy entities. To meet these requirements, previous studies proposed to automatically construct a corpus from dictionaries (Noraset et al., 2017; Gadetsky et al., 2018) and online encyclopedias (Ni and Wang, 2017; Ishiwatari et al., 2019). These corpora have significantly contributed to the development of definition modelling techniques (Mickus et al., 2019; Washio et al., 2019; Li et al., 2020, to name a few). However, despite its importance in critical applications, such a corpus is scarce in languages other than English.

This study created a corpus for definition modelling in Japanese, JADE (corpus for JApanese DEfinition modelling) using Wikipedia[1] and Wikidata[2] following Ishiwatari et al. (2019). Wikipedia has two advantages with regard to creating a corpus for definition modelling. The first is the coverage as Wikipedia covers a variety of topics in a timely fashion, and the second is its multilinguality in that datasets can easily be created in languages other than English, *i.e.* Japanese. The JADE corpus is sufficiently large to train neural models providing about 630k set of target words/phrases, their definitions, and contexts, for about 41k unique targets. The performance of the state-of-the-art definition modelling method proposed by Huang et al. (2021) is also benchmarked. The JADE corpus[3] was released for future research on Japanese definition modelling.

2. Related Work

Corpora for definition modelling have primarily been created for English. Noraset et al. (2017) created the Wordnet dataset by collecting entries in the GNU Collaborative International Dictionary of English[4] and Wordnet glosses (Miller, 1995) and the original dataset provides only a target and its definition. Later Ishiwatari et al. (2019) expanded this dataset by adding contexts. While the Wordnet dataset targets on general words, Gadetsky et al. (2018) created the Oxford dataset targeting on polysemous words. They collected entries using the APIs of Oxford Dictionaries.[5]

The Wordnet and Oxford datasets were both created from common dictionaries; hence, they do not pro-

[†] This study was conducted when the first author was an graduate student at Osaka University.

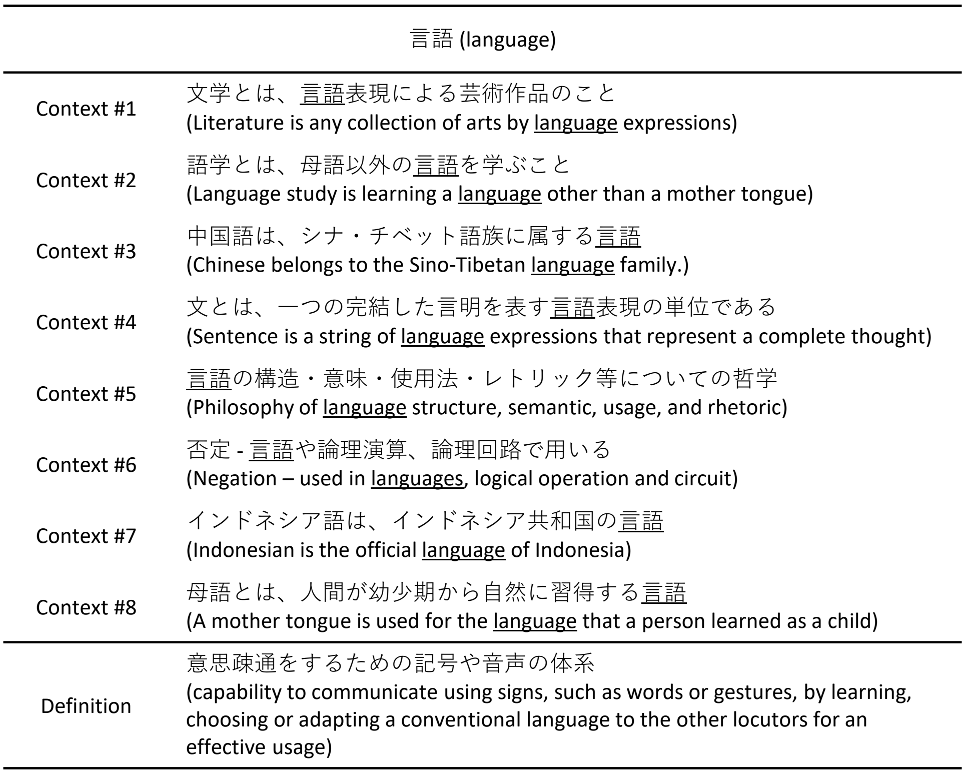
[1] https://en.wikipedia.org/wiki/Main_Page

[2] https://www.wikidata.org/wiki/Wikidata:Main_Page

[3] https://doi.org/10.5281/zenodo.5513039

[4] http://wwwgcide.gnu.org.ua

[5] https://developer.oxforddictionaries.com/



言語 (language)	
Context #1	文学とは、言語表現による芸術作品のこと (Literature is any collection of arts by language expressions)
Context #2	語学とは、母語以外の言語を学ぶこと (Language study is learning a language other than a mother tongue)
Context #3	中国語は、シナ・チベット語族に属する言語 (Chinese belongs to the Sino-Tibetan language family.)
Context #4	文とは、一つの完結した言明を表す言語表現の単位である (Sentence is a string of language expressions that represent a complete thought)
Context #5	言語の構造・意味・使用法・レトリック等についての哲学 (Philosophy of language structure, semantic, usage, and rhetoric)
Context #6	否定 - 言語や論理演算、論理回路で用いる (Negation – used in languages, logical operation and circuit)
Context #7	インドネシア語は、インドネシア共和国の言語 (Indonesian is the official language of Indonesia)
Context #8	母語とは、人間が幼少期から自然に習得する言語 (A mother tongue is used for the language that a person learned as a child)
Definition	意思疎通をするための記号や音声の体系 (capability to communicate using signs, such as words or gestures, by learning, choosing or adapting a conventional language to the other locutors for an effective usage)

Figure 1: Examples of entries in the JADE corpus (English translation is in parentheses, wherein the English definition is extracted from the original Wikidata item.)

vide neologisms and rare expressions. To complement these expressions, Ni and Wang (2017) created the Urban dataset using the Urban Dictionary[6], which is the largest online slang dictionary where definitions and examples are submitted by internet users. Unlike the Wordnet and Oxford datasets, the Urban dataset also provides phrases as targets.

HEI++ (Bevilacqua et al., 2020) targets on free phrases (*e.g.* 'exotic cuisine') that are rarely included in common dictionaries. HEI++ provides high quality definitions written by an expert lexicographer for 713 adjective-noun phrases, although the scale is limited.

Ishiwatari et al. (2019) created a largest ever corpus for definition modelling that covers both words and phrases using Wikipedia and Wikidata. While the Urban dataset also covers phrases, its domain is limited to internet slangs. In contrast, the Wikipedia dataset covers phrases in a variety of domains and topics. JADE is direct successor of this Wikipedia dataset and the first corpus for Japanese definition modelling.

3. Construction of JADE Corpus

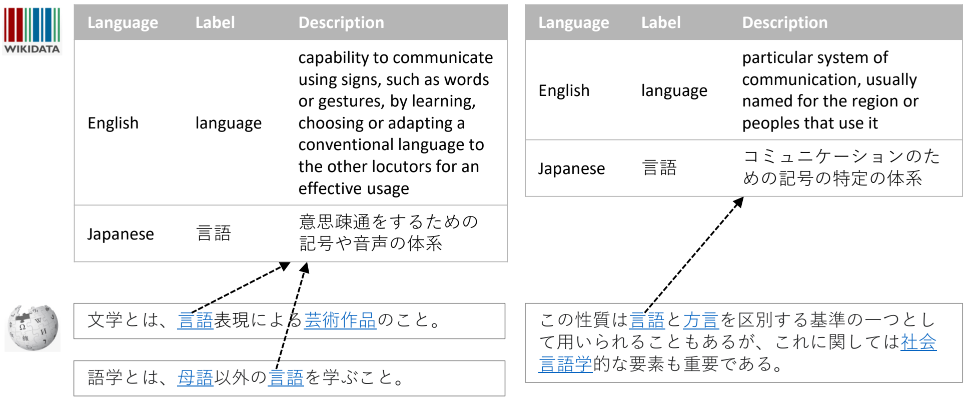
Each entry in JADE consists of (1) a word or phrase to be defined, which is called *target* hereinafter, (2) its definition, and (3) a context (*i.e.* an usage example) containing the target. Given the nature of Wikipedia, most target words and phrases of JADE are proper nouns. Figure 1 shows examples of entries in JADE, whose target is 'language' defined as 'capability to communicate using signs, such as words or gestures, by learning, choosing or adapting a conventional language to the other locutors for an effective usage.' The target has eight contexts in total, each of which shows how the target is used in a sentence.

Following Ishiwatari et al. (2019), JADE was created by mining Japanese Wikipedia and Wikidata, as illustrated in Figure 2. Wikidata acts as the central storage for the structured data of Wikipedia, which provides definitions of topics, concepts, and objects in different languages. Targets and their definitions were first extracted in Japanese from the Wikidata items. Next, to collect contexts of targets, sentences from the first paragraphs of Wikipedia articles that have links to other Wikipedia pages were crawled and finally, a target and its context referring the linked article were matched where the corresponding Wikidata item is associated with. Note that this study only used links whose link text exactly matches with a title of Wikipedia article so that links attached just for introducing related articles can be excluded.

This process allowed for the ambiguity in meanings of polysemous words and phrases to be resolved without human intervention. Note that a target may have multiple entries, *i.e.* multiple entries may have the same target, when a target is polysemous and/or the target has multiple contexts.

We collected $629,637$ sets of targets, definitions, and contexts in total, and split them into train, validation,

[6] https://www.urbandictionary.com/



Language	Label	Description
English	language	capability to communicate using signs, such as words or gestures, by learning, choosing or adapting a conventional language to the other locutors for an effective usage
Japanese	言語	意思疎通をするための記号や音声の体系

Language	Label	Description
English	language	particular system of communication, usually named for the region or peoples that use it
Japanese	言語	コミュニケーションのための記号の特定の体系

文学とは、言語表現による芸術作品のこと。

語学とは、母語以外の言語を学ぶこと。

この性質は言語と方言を区別する基準の一つとして用いられることもあるが、これに関しては社会言語学的な要素も重要である。

Figure 2: JADE creation process; linking Wikidata items with sentences in Wikipedia articles

Split	Number of unique targets	Number of entries	Context length	Definition length
Train	$36,568$	$574,188$	17.1	11.6
Valid	$2,041$	$25,068$	17.1	10.8
Test	$1,956$	$30,381$	16.7	11.0
ALL	$40,565$	$629,637$	17.0	11.1

Table 1: Statistics of the JADE corpus ('length' indicates the average number of tokens in a sentence)

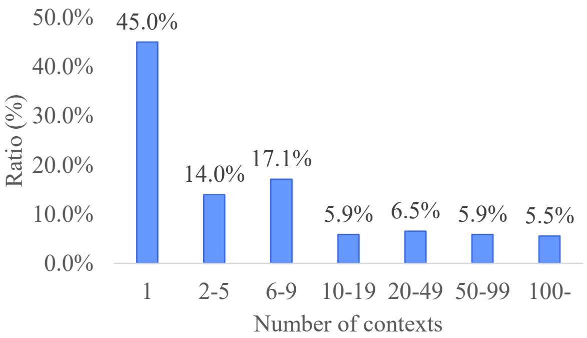


Figure 3: Distribution of numbers of contexts in JADE

and test sets, ensuring that the same target does not appear in both training and validation/test sets. The statistics of the JADE corpus are shown in Table 1, where definitions and contexts have 11.1 and 17.0 tokens on average, respectively.[7] Figure 3 shows the distribution of numbers of contexts per definition, indicating that 55.0% of definitions in the JADE corpus has multiple contexts.

4. Benchmark

The state-of-the-art definition modelling method was evaluated on the JADE corpus for future development of the technique in Japanese. The BLEU (Papineni et al., 2002) and NIST (Doddington, 2002) scores were reported following the previous studies. NIST focuses on content words by giving more weights to them, which makes NIST more informative than solely assigning an equal weight to each m-gram as in BLEU.[8] This caracteristic is crucial on definition generation because references are shorter and tend to follow a template. For example, a typical template is 'X is Y of Z,' such as 'Paris is the capital of France.'

4.1. Models

The performance of a definition modelling method proposed by Huang et al. (2021) was evaluated. Huang et al. (2021) rerank n-best definitions generated by a fine-tuned Text-to-Text Transfer Transformer (T5) model (Raffel et al., 2020) to control the level of specificity. Namely, they employ two evaluators: one evaluates the level of over-specificity of a definition and the other evaluates the level of under-specificity.

They fine-tune additional T5 models and use force-decoding for estimating the levels of over/under-specificity. For the former, they force-decode a context conditioned on a definition, assuming that an over-specific definition likely generates the context because it tends to be overly affected by the context. For the latter, they force-decode a definition conditioned only on a target, assuming that an under-specific definition can be easily force-decoded from the target because it likely captures the most general meaning of the target. While their paper reported evaluation results using English corpora discussed in Section 2, the method can be easily extended to a multilingual setting by replacing T5 with its multilingual version (mT5) (Xue et al., 2021).

[7] We used Stanza (Qi et al., 2020) for tokenization.

[8] We computed a NIST score using the NLTK library: `https://www.nltk.org/_modules/nltk/translate/nist_score.html`

Target	回路図 (circuit diagram)
Context	回路図とレイアウトの照合を行うソフトウェア (A software that determines whether a particular circuit layout corresponds to the circuit diagram)
Reference	回路を記述するために用いられる図 (A graphical representation to describe an electrical circuit)
mT5-base	電磁気学における諸方程式 (A series of equations of electromagnetism)
(Huang et al. 2021)	電磁気学において、回路の上の制御に関与する地図 (A map concerning a control on a circuit in electromagnetism)
Target	王族 (royal house)
Context	伯邑考は、周の王族 (Bo Yikao was a member of the Zhou royal house)
Reference	国王の家族・親族 (Family and relatives of a monarch)
mT5-base	共通の祖先を持つ血縁集団 (A descent group having the common ancestor)
(Huang et al. 2021)	国王の家族・親族 (Family and relatives of a monarch)
Target	人工衛星 (artificial satellite)
Context	人工衛星の開発なども行っていた ((They) also developed artificial satellites)
Reference	地球を焦点の1つとする楕円軌道を周回する人工の天体 (A celestial body created artificially that goes around an elliptical orbit having the earth as its focus)
mT5-base	太陽系を含んだ銀河 (A galaxy including the solar system)
(Huang et al. 2021)	人工的な手段を用いて意図的に作られた衛星 (A satellite created by artificial means with certain purposes)

Figure 4: Samples of generated definitions (English translations are in parenthesis)

Besides, the performance of an mT5 fine-tuned on JADE was also measured. Huang et al. (2021) reported that this simple baseline already outperformed previous methods for definition modelling with a large margin on English tasks.

Implementation Details For both methods, the mT5-base model released at HuggingFace[9] was used and conducted fine-tuning using the training set of the JADE corpus. As for the method by Huang et al. (2021), the hyper-parameters of α and β were tuned to maximise BLEU and NIST scores, respectively, using the development set of JADE. Specifically, we set $(\alpha, \beta) = (0.6, 0.2)$ for BLEU and $(\alpha, \beta) = (0.6, 0.1)$ for NIST. At inference, 100 definitions per target were generated and reranked.

4.2. Results

Table 2 shows BLEU and NIST scores measured on the test set of JADE corpus with the results of statistical significance testing. We used the Wilcoxon signed-rank test (Wilcoxon, 1945), which tests the null hypothesis that two related paired samples are from the same distribution.

	BLEU	NIST
mT5-base	44.59	154.17
Huang et al. (2021)	**46.06***	**161.57***

Table 2: BLEU and NIST scores measured on the JADE test set (* indicates a significant difference at $p < 0.05$)

The results indicate that the method proposed by Huang et al. (2021) significantly outperformed mT5. Figure 4 shows example definitions generated by these methods. Unsurprisingly, both methods generated highly fluent definitions thanks to mT5's pre-training using huge corpora. However, mT5 generated incorrect definitions of targets for the first and third examples. As for the second example, mT5 generated a too general

[9] https://huggingface.co/google/mt5-base

definition that lacks meaning of 'royal.' In contrast, Huang et al. (2021) generated an appropriate definition for this target. As for the third example, while the generated definition by Huang et al. (2021) describes a different aspect of 'artificial satellite,' the definition is acceptable from human perspectives.

5. Summary

This study details the development and creation of the JADE corpus, a corpus for Japanese definition modelling, by matching Wikidata items and Wikipedia articles. Thanks to the nature of the data-source, JADE provides words and phrases in variety of domains and topics as targets. The size of JADE corpus allows not only testing but also training of neural network based models.

The JADE corpus was released online for future research on definition modelling for the Japanese language. Future work should include complementing the JADE corpus to cover free phrases as the HEI++ corpus does. Such free phrases would contribute to representation learning for lexical semantics, particularly for non-compositional phrases. Besides, we can create a cross-lingual definition modelling dataset. Cross-lingual definition generation should be useful for supporting language learners.

Acknowledgement

This work was supported by JSPS KAKENHI Grant Number 21H03564.

Bibliographical References

Bevilacqua, M., Maru, M., and Navigli, R. (2020). Generationary or "how we went beyond word sense inventories and learned to gloss". In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7207–7221, November.

Doddington, G. (2002). Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the Human Language Technology Conference (HLT)*, pages 138––145.

Gadetsky, A., Yakubovskiy, I., and Vetrov, D. (2018). Conditional generators of words definitions. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 266–271, July.

Huang, H., Kajiwara, T., and Arase, Y. (2021). Definition modelling for appropriate specificity. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2499–2509, November.

Ishiwatari, S., Hayashi, H., Yoshinaga, N., Neubig, G., Sato, S., Toyoda, M., and Kitsuregawa, M. (2019). Learning to describe unknown phrases with local and global contexts. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 3467–3476, June.

Li, J., Bao, Y., Huang, S., Dai, X., and Chen, J. (2020). Explicit semantic decomposition for definition generation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 708–717, July.

Mickus, T., Paperno, D., and Constant, M. (2019). Mark my word: A sequence-to-sequence approach to definition modeling. In *Proceedings of the NLPL Workshop on Deep Learning for Natural Language Processing*, pages 1–11, September.

Miller, G. A. (1995). WordNet: A lexical database for english. *Communications of the ACM*, 38(11):39––41, November.

Ni, K. and Wang, W. Y. (2017). Learning to explain non-standard English words and phrases. In *Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP)*, pages 413–417, November.

Noraset, T., Liang, C., Birnbaum, L., and Downey, D. (2017). Definition modeling: Learning to define word embeddings in natural language. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 3259–3266.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 311–318, July.

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., and Manning, C. D. (2020). Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 101–108, July.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Washio, K., Sekine, S., and Kato, T. (2019). Bridging the defined and the defining: Exploiting implicit lexical semantic relations in definition modeling. In *Proceedings of Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3521–3527, November.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83.

Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., and Raffel, C. (2021). mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 483–498, June.