Utilizing Natural Language Processing Technologies for Controlled Lexicon Building: A Pilot Study Focusing on English and Japanese Verbs

Daichi Yamaguchi[1], Hodai Sugino[1], Rei Miyata[2], Satoshi Sato[1]
[1]Nagoya University, [2]The University of Tokyo, Japan
E-mail: yamaguchi.daichi.e4@s.mail.nagoya-u.ac.jp, 89sugino1230@gmail.com,
ray.miyata@gmail.com, sato.satoshi.g9@f.mail.nagoya-u.ac.jp

Keywords: controlled lexicon building; word variation management; interchangeability of words; natural language processing application; automotive domain

This paper presents our ongoing research project to automate the process of controlled lexicon building.

A controlled lexicon is a set of approved words defined for a specific purpose, such as controlled authoring and translation (ASD, 2021; Møller & Christoffersen, 2006; Warburton, 2014). The proper use of a controlled lexicon can prevent textual variation, leading to improved text consistency and clarity. Although many controlled lexicons have been built for various purposes (Kuhn, 2014), there have been few examinations of the possibility of automating lexicon creation. Fundamentally, the process of building a controlled lexicon has not been well formalized. Miyata & Sugino (2020) presented corpus-based lexicon-building procedures and proposed the *interchangeability* of words in actual sentences as a key criterion to identify word variations. Nevertheless, their lexicon-building process mostly depends on human expertise, and the detailed steps for judging interchangeability have yet to be clarified. Because natural language processing (NLP) technologies based on deep learning have advanced rapidly, we envisage the effective use of such technologies in this process.

Hence, towards the automation of controlled lexicon building, we have formalized the lexicon-building process and examined the applicability of various NLP technologies. Following the corpus-based procedures in (Miyata & Sugino, 2020), the process of controlled lexicon building can be broadly divided into the following two steps:

Step 1. Connect words that are interchangeable to form word clusters.
Step 2. For each cluster, define one word as approved and the rest as unapproved.

In Step 1, to capture interchangeability, we quantify the different levels of word similarity using various NLP technologies:

(a) General similarity: Word embeddings, such as word2vec (Mikolov et al., 2013), trained on general domain corpora, such as web text, can be used. Conventional thesauri, such as WordNet (Princeton University, 2010), can also be used.
(b) Domain-specific similarity: Word embeddings trained on target domain corpora can be used.
(c) Domain-specific context-aware similarity: Contextualized embedding methods, such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019), can be used.

For level (c), we examined the interchangeability of words in example sentences in the target domain corpus. For example, the verb “delete” can be replaced with “erase” in an example sentence “Delete the data”. If this consistently applies to other example sentences in

the target corpus, we can assume that a controlled lexicon should include either word but not both to avoid variation. To simulate human judgments regarding interchangeability, we used contextualized embedding methods to produce vector representations that encode not only target words but also their context.

In Step 2, we tested several algorithms to define the approved words based on the word frequency and the linguistic symmetricity of their antonyms. Although the frequency of words in the target corpus can be regarded as a major factor in deciding the approved words, the symmetricity of certain word pairs in a lexicon can sometimes precede frequency evidence. For example, if the verb "engage" is already defined as approved, the symmetric verb "disengage" is likely to be selected as approved instead of a synonymous verb "detach", even if the latter is more frequently observed in the corpus than the former. To capture the symmetricity of words, we devised language-specific heuristic rules that use linguistic or textual clues, such as verb constructions (e.g., *sa*-hen noun + *suru* construction) and n-gram overlap at the character level (e.g., "**engage**" and "dis**engage**").

First, we explain our overall framework for automating the process of controlled lexicon building. We then present the results of our pilot experiments applying various NLP technologies to each lexicon-building step, focusing on English and Japanese verbs observed in automotive domain corpora. The obtained lists of approved and unapproved words are next compared with a controlled lexicon manually constructed from the same corpora. These results suggest to what extent current technologies can help with specialized lexicographic work.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Numbers 19H05660 and 23H03689. The automobile manuals used in this study were provided by Toyota Motor Corporation.

References

- ASD (2021). ASD Simplified Technical English. Specification ASD-STE100, Issue 8. URL: http://www.asd-ste100.org.

- Devlin, J., Chang, M.W., Lee, K., 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. Minneapolis, Minnesota, USA, pp. 4171–4186.

- Kuhn, T. (2014). A Survey and Classification of Controlled Natural Languages. Computational Linguistics, 40(1), pp. 121–170.

- Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. CoRR, abs/1907.11692.

- Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. In C. Burges, L. Bottou, M. Welling, Z. Ghahramani & K. Weinberger (eds.) Advances in Neural Information Processing Systems 26. Curran Associates, Inc., pp. 3111–3119.
- Miyata, R., Sugino, H. (2020). Building a Controlled Lexicon for Authoring Automotive Technical Documents. In Z. Gavriilidou, M. Mitsiaki & A. Fliatouras (eds.) Lexicography for Inclusion: Proceedings of the 19th EURALEX International Congress, 7-9 September 2021, Alexandroupolis, Vol. 1. Democritus University of Thrace, pp. 171–180.
- Møller, M.H., Christoffersen, E. (2006). Buiding a Controlled Language Lexicon for Danish. LSP & Professional Communication, 6(1), pp. 26–37.
- Princeton University (2010). About WordNet. WordNet. URL: https://wordnet.princeton.edu/.
- Warburton, K. (2014). Developing Lexical Resources for Controlled Authoring Purposes. In Proceedings of LREC 2014 Workshop: Controlled Natural Language Simplifying Language Use. Reykjavik, Iceland, pp. 90–103.