Robust Automatic Annotation of Argument Structure Constructions

Kristopher Kyle, Department of Linguistics, University of Oregon, kkyle2@uoregon.edu
Hakyung Sung, Department of Linguistics, University of Oregon, hsung@uoregon.edu

Argument structure constructions (ASC) are commonly extracted from corpora for a range of research purposes in Linguistics. These include the investigation alternation (e.g., Gries & Wulff, 2009), the analysis of verb-construction contingencies (e.g., Ellis & Ferreira-Junior, 2009a, b; Kyle & Crossley, 2017), examining the constructicon of first and second language users (e.g., Römer et al., 2014), and the measurement of proficiency/development (e.g., Hwang & Kim, 2022). An important issue in studies that analyze the characteristics of ASC use is the method used to identify ASCs and their verbs. Many studies have used a manual approach to identify ASCs in relatively small corpora (e.g., Goldberg et al., 2004; Ellis & Ferreira-Junior, 2009a, b). Given the increase in the availability of large datasets of learner data, automatic methods of ASC extraction have been proposed, including the use of syntactic frames as ASCs (e.g., Kyle & Crossley, 2017) and rule-based systems that rely on syntactic frames and explicit lexical information (Hwang & Kim, 2022). To date, however, no approach has used machine- learning techniques to predict ASCs directly, primarily because no ASC treebank has been available.

In this study, we first developed an ASC treebank by building on the English portion of the Universal Propositions (UP) project (Akbik et al., 2015), which represents a merge of the Universal Dependencies version of the English Web Treebank (EWT; Bies et al., 2012; Silveria et al., 2014) and PropBank (Palmer et al., 2005). For each sentence in the training section of UP, we extracted the large-grained argument structures (e.g., *ARG0-Verbsense-ARG1*) and converted them to fine-grained semantic role frames (e.g., *agent-V-theme*) using relation mappings from PropBank. We then manually assigned an ASC to each semantic role frame that occurred at least five times in the corpus. In total, 26,437 ASC instances were annotated and included in the analysis with nine representative ASC tags (i.e., attributive, caused-motion, ditransitive, intransitive-simple, intransitive-motion, intransitive-resultative, passive, transitive-simple, and transitive-resultative). The ASC Treebank will be made publicly available to the research community.

Based on the ASC Treebank, we trained three probabilistic models that relied on varying linguistic information (main verb lemma, syntactic frame, main verb lemma + syntactic frame) and a multiclass transformer model based on RoBERTa (Liu, 2019) embeddings. The results indicated that the transformer model achieved the highest overall classification accuracy (F1 = .918), followed by the verb lemma+syntactic frame model, the syntactic frame model, and the verb lemma model. With regard to individual ASC types, the transformer model also achieved the highest F1 score for each of the nine ASCs represented in the treebank. Individual annotation accuracy scores for the transformer model ranged from F1 = .982 for attributive constructions to F1 =. 742 for caused motion constructions (which had relatively low representation in the training data).

In this presentation, we discuss the strengths and weaknesses of each model as they pertain to both practical and theoretical concerns. We also outline future directions for the expansion and further development of the ASC Treebank and the automatic annotation tool.

References

Akbik, A., Chiticariu, L., Danilevsky, M., Li, Y., Vaithyanathan, S., & Zhu, H. (2015, July). Generating high quality proposition banks for multilingual semantic role labeling. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1),* 397-407.

Bies, A., Mott, J., Warner, C., & Kulick, S. (2012). English web treebank. *Linguistic Data Consortium, Philadelphia, PA*.

Ellis, N. C., & Ferreira-Junior, F. (2009a). Constructions and their acquisition: Islands and the distinctiveness of their occupancy. *Annual Review of Cognitive Linguistics*, *7*(1), 188-221.

Ellis, N. C., & Ferreira–Junior, F. (2009b). Construction learning as a function of frequency, frequency distribution, and function. *The Modern language journal*, *93*(3), 370-385.

Goldberg, A. E., Casenhiser, D. M., & Sethuraman, N. (2004). Learning argument structure generalizations. *Cognitive Linguistics*, *15*(3), 289-316.

Gries, S. T., & Wulff, S. (2009). Psycholinguistic and corpus-linguistic evidence for L2 constructions. *Annual Review of Cognitive Linguistics*, *7*(1), 163-186.

Hwang, H., & Kim, H. (2022). Automatic Analysis of Constructional Diversity as a Predictor of EFL Students' Writing Proficiency. *Applied Linguistics*.

Kyle, K., & Crossley, S. (2017). Assessing syntactic sophistication in L2 writing: A usage-based approach. *Language Testing*, *34*(4), 513-535.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Palmer, M., Gildea, D., & Kingsbury, P. (2005). The proposition bank: An annotated corpus of semantic roles. *Computational linguistics*, *31*(1), 71-106.

Römer, U., Roberson, A., O'Donnell, M. B., & Ellis, N. C. (2014). Linking learner corpus and experimental data in studying second language learners' knowledge of verb-argument constructions. *ICAME Journal*, *38*(1), 115–135.

Silveira, N., Dozat, T., De Marneffe, M.-C., Bowman, S., Connor, M., Bauer, J., & Manning, C. (2014). A Gold Standard Dependency Corpus for English. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, 2897–2904.