Reforging : A Method for Constructing a Linguistically Valid Japanese CCG Treebank

Asa Tomita [1] Hitomi Yanaka [2] Daisuke Bekki [1]
Ochanomizu University [1] The University of Tokyo [2]
tomita.asa@is.ocha.ac.jp

EACL 2024

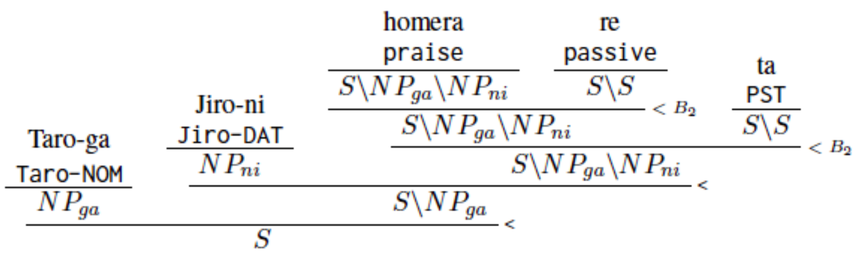
Abstract

- The linguistic validity of Combinatory Categorial Grammar (CCG; Steedman, 1996, 2000) parsing results relies on the validity of treebanks for training and evaluation data, so the treebank construction is crucial
- We propose a method to generate a linguistically valid Japanese CCG treebank with detailed information by combining the strengths of ABCTreebank and Japanese CCG parser *lightblue.*
- We manually evaluate CCG syntactic structures and semantic representations, and syntactically and semantically valid trees were produced for 19 of 56 sentences (33%).

Background

Japanese CCGbank [Uematsu et al., 2013]
- was constructed by automatic conversion of Japanese dependency tree corpora.

✗ It contains some empirical fallacies on predictions for passive and causative nestings.



ABCTreebank [Kubota et al., 2020]
- was constructed by converting the Keyaki Treebank (phrase structured treebank) to ABC grammar tree.

✓ Argument structures are assumed to be reliable because they were manually annotated.

✗ It does not cover the syntactic information, such as syntactic features.

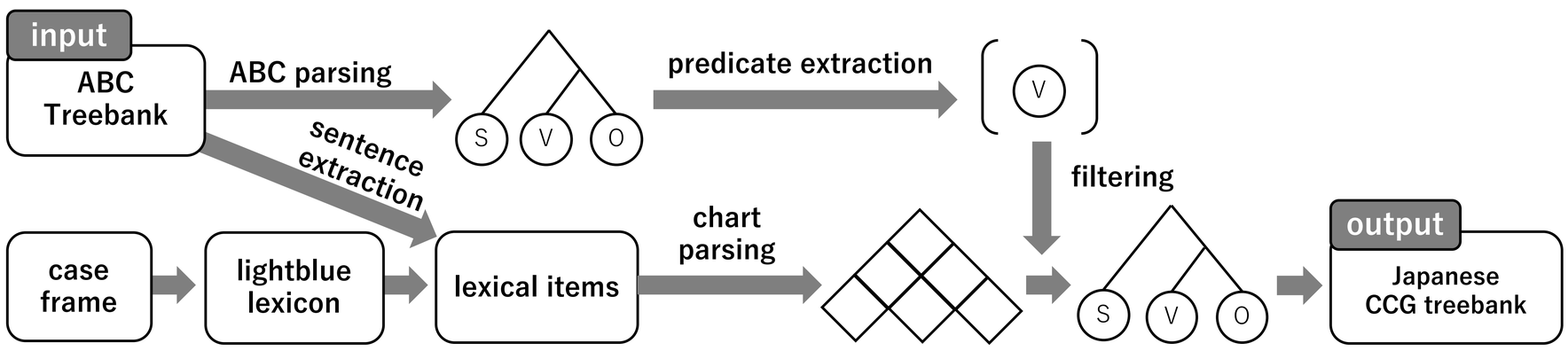
lightblue [bekki,2010]
- is a Japanese CCG parser that computes syntactic structures from lexicon and combinatory rules, so it does not require training and evaluation data.

✓ It outputs CCG syntactic structures with detail syntactic feature and semantic representations in terms of Dependent Type Semantics (DTS; Bekki and Mineshima, 2017)

✗ It includes some argument structure errors.

The Reforging Process and Treebank Construction

We proposed a method for constructing a Japanese CCG treebank by combining the positive aspects of ABCTreebank, in which argument structures are manually annotated, with lightblue's abilty to provide CCG trees with detailed syntactic features.



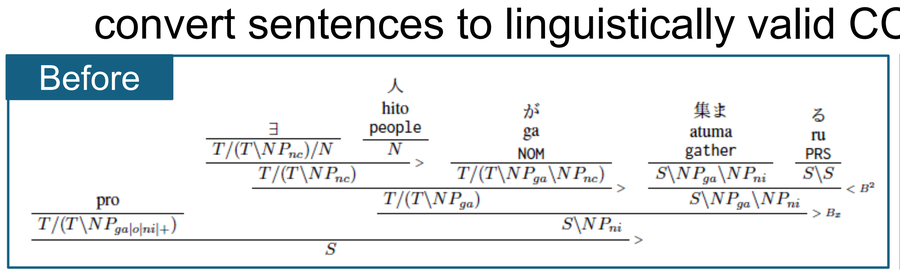
Genre	Sentences	Len-50+ sentences	Reforged trees
aozora	1773	590	1183
bible	1652	220	1430
book_expert	50	4	41
dict_lexicon	2640	4	2636
diet_kaigiroku	486	112	374
fiction	921	44	877
law	337	128	209
misc	335	59	276
news	443	103	340
non-fiction	223	87	126
spoken	570	11	559
ted_talk	605	54	551
text-book	4880	10	4870
wikipedia	222	51	171
Total	15137	1482	13653

Statistics for the reforged treebank data.

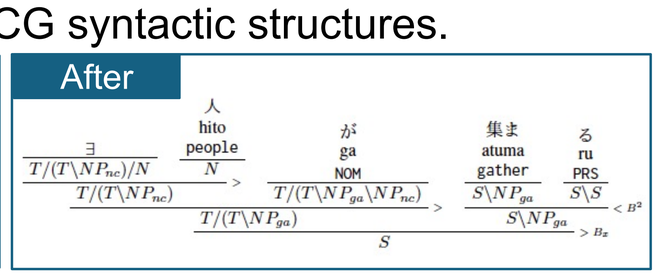
Discussion

- The trees below show the syntactic structures of the sentence "人が集まる" (People-NOM gather) before and after *reforging*.
  - *gather* was analyzed as having ga-case (nominative case) and ni-case (dative case) argument positions before reforging.
  - By overwriting the syntactic structure of *gather* with a lexical entry having only ga-case (nominative case) argument position, it became possible to convert sentences to linguistically valid CCG syntactic structures.

Before



After



Evaluation

- We manually evaluated 56 randomly sampled sentences, four from each genre.
- Syntactically and semantically valid trees were produced for 19 sentences (33%).

	Metrics	Sentences
Syntactic Error	Syntactic category	18
	Compound verb	4
	Other syntactic error	30
Semantic Error		7
No Error		19

Error Analysis

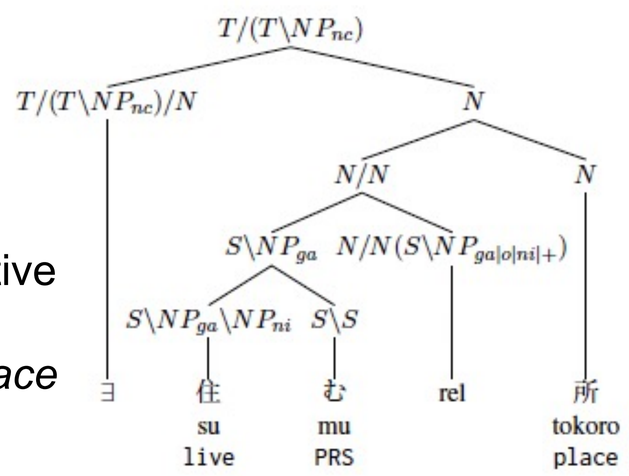
Sentence example involving a relative clause :

"食べる ものも なけれ ば、住む ところも ない"

Eat-attr thing-NOM not-exist and, live-attr place-NOM not-exist

'No food to eat, no place to live'

- The predicate *live* takes the ga-case (nominative case) NP and becomes a relative clause.
  → An external relation between *live* and *place* was analyzed as an internal relation.



Conclusion

- We proposed a reforging method for constructing linguistically valid Japanese CCG treebank with detailed syntactic features.
- Our future work will consider combining ABCTreebank with other reliable resources, and improve *lightblue*'s parsing algorithm to better handle long sentences.