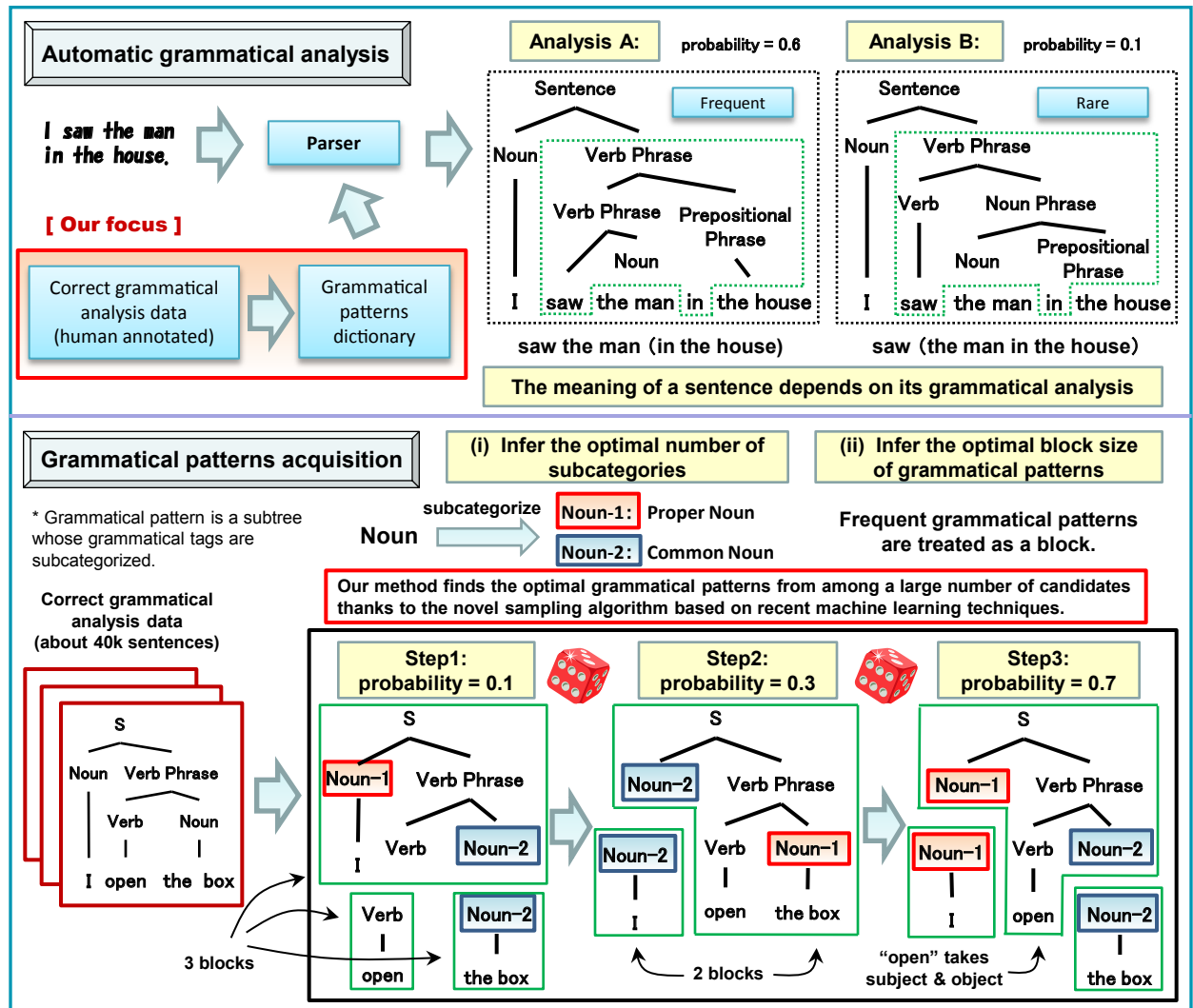# 12 Automatic grammatical analysis of English

Syntactic parsing based on statistical grammar induction

Abstract— We need to grasp the subject and predicate of sentences to understand English, Japanese and other languages. However, it is not easy for a computer to perform such grammatical analysis automatically. We have developed grammatical analysis software based on the statistical acquisition of grammatical patterns from English resources. We employed this approach for parsing and achieved the best performance on international benchmark data. In the future, our aim is to enable computers to understand language and thus support human communication activities. As the first step towards our goal, we expect the grammatical analyzer to be useful for achieving high-quality machine translation and automatic language correction systems.

Automatic grammatical analysis

I saw the man in the house. → Parser →

[Our focus]

Correct grammatical analysis data (human annotated) → Grammatical patterns dictionary

Analysis A: probability = 0.6 — Frequent

Sentence / Noun / Verb Phrase / Verb Phrase / Prepositional Phrase / Noun
I saw the man in the house
saw the man (in the house)

Analysis B: probability = 0.1 — Rare

Sentence / Noun / Verb Phrase / Verb / Noun Phrase / Noun / Prepositional Phrase
I saw the man in the house
saw (the man in the house)

The meaning of a sentence depends on its grammatical analysis

Grammatical patterns acquisition

* Grammatical pattern is a subtree whose grammatical tags are subcategorized.

(i) Infer the optimal number of subcategories

Noun → subcategorize → Noun-1: Proper Noun / Noun-2: Common Noun

(ii) Infer the optimal block size of grammatical patterns

Frequent grammatical patterns are treated as a block.

Our method finds the optimal grammatical patterns from among a large number of candidates thanks to the novel sampling algorithm based on recent machine learning techniques.

Correct grammatical analysis data (about 40k sentences)

S / Noun / Verb Phrase / Verb / Noun
I open the box

Step1: probability = 0.1
S / Noun-1 / Verb Phrase / Verb / Noun-2
I
Verb: open / Noun-2: the box
3 blocks

Step2: probability = 0.3
S / Noun-2 / Verb Phrase / Verb / Noun-1
Noun-2: I
open the box
2 blocks

Step3: probability = 0.7
S / Noun-1 / Verb Phrase / Noun-1 / Verb / Noun-2
I / open / Noun-2: the box
"open" takes subject & object

Related works

[1] H. Shindo, Y. Miyao, A. Fujino, M. Nagata, "Bayesian Symbol-Refined Tree Substitution Grammars for Syntactic Parsing," in *Proc. the annual meeting of the Association for Computational Linguistics (ACL)*, 2012 (Best Paper Award).
[2] H. Shindo, A. Fujino, M. Nagata, "Insertion Operator for Bayesian Tree Substitution Grammars," in *Proc. the annual meeting of the Association for Computational Linguistics (ACL)*, 2011.

Contact

Hiroyuki Shindo Intelligence Research Group, Innovation Communication Laboratory
E-mail : shindo.hiroyuki{at}lab.ntt.co.jp (Please replace {at} with @)