COMBINING STATISTICAL MODELS WITH SYMBOLIC GRAMMAR IN PARSING

Junichi Tsujii

Department of Computer Science, University of Tokyo, Japan
School of Computer Science, University of Manchester, UK

ABSTRACT

There are two streams of research in Computational Linguistics and Natural language Processing, the empiricist and rationalist traditions. Theories and computational techniques in these two streams have been developed separately and different in nature. Although the two traditions have been considered irreconcilable and have often been antagonistic toward each other, I have contention with this assertion, and thus claim that these two research streams in linguistics, despite or due to their differences, can be complementary to each other and should be combined into a unified methodology. I will demonstrate in my talk that there have been interesting developments in this direction of integration, and would like to discuss some of the recent results with their implications on engineering application.

Computational Linguistics (CL), Natural Language Processing (NLP), and their engineering application have made a significant amount of progress in the last decade. The recent progress has largely been indebted to the progress made in techniques for the statistical modeling of language, which has been inspired by and is closely linked with the technologies developed by the speech community.

Statistical language modeling fits well with the empiricist tradition in linguistics, or with the corpus-based linguistics, which emphasizes the importance of observable data in constructing linguistic theories. The extreme perspective of this position is that no prior theoretical concepts should be introduced, and that every concept in the theory should be inductively derived from data.

On the other hand, there exists another tradition in linguistics: the rationalist tradition, or theoretical linguistics. Computational linguistics was formerly taken as a branch of this tradition with a particular emphasis on formal aspects. Although this tradition is also interested in the "modeling of language", it sees language as a complex system that follows a set of rules. Instead of using inductive methods in theory construction, this tradition emphasizes the deductive aspect of a theory, and consequently claims that a theory should be able to deduce a set of statements. Statements thus derived from a theory should be able to be checked against data in order to validate or refute the theory.

In its simplest form, a model of language is described by a set of rules which generates an infinite set of word sequences (generative definition of language). If a given set of rules generates the set of all word sequences, and only word sequences that belong to a language, the set of rules or grammar is said to be adequate as a description (or model) of the language.

Further extended models of language in this tradition postulate different layers of representation of a sentence, such as;

1. Surface sequence of words
2. Constituent structure
3. Semantic structure
4. Contextual structure,

and they formulate grammar as constraints that exist among representations of such layers. A linguistic object, which consists of these layers of representation, should satisfy all constraints for it to be legitimate in the language.

Since all layers of representation, except for surface sequences of words, are unobservable, and thus have to be postulated a priori to observable data, such models cannot be constructed using purely empiricist approaches. Furthermore, reflecting the emphasis on formal aspects, computational linguists have devoted themselves to develop a mathematically well-defined framework for describing constraints, so that mechanical means can manipulate descriptions. Such formal frameworks are called grammar formalisms.

Though there still remain differences among different grammar formalisms such as HPSG (Head-driven Phrase Structure Grammar), LFG (Lexical Functional Grammar), CCG (Combinatorial Categorical Grammar), TAG (Tree Adjoining Grammar), etc., efforts by computational linguists have culminated in frameworks collectively called Unification-based formalisms, which use complex feature structures to describe constraints on linguistic objects.

978-1-4244-1746-9/07/$25.00 ©2007 IEEE

Furthermore, I would like to demonstrate that there have been interesting developments in this direction of integration, and would like to discuss some of the recent results with their implications on engineering application.
The issues I would like to address in my talk are:

(1) Merits and Difficulties in Rationalist Grammar
Why the semantic layer of representation is essential for NLP and important for speech recognition.
The impossibility of developing a rationalist grammar and making it work.
(2) Corpus-based Grammar Development
How we combine the two paradigms of linguistics to produce a feasible engineering framework for grammar development.
(3) Preferences and Constraints
The fragility of rationalist grammar
How we can combine a rationalist grammar with statistical modeling of language to make it robust.
(4) Efficient Parsing
The importance of mathematically sound formalisms for making parsing efficient
Decomposition of constraints
How we increase the efficiency of make parsing based on grammar formalisms.
(5) New Paradigms of Parsing