Things between Lexicon and Grammar (Extended Abstract)

Yuji Matsumoto
Graduate School of Information Science
Nara Institute of Science and Technology (NAIST)
matsu@is.naist.jp

A number of grammar formalisms were proposed in 80's, such as Lexical Functional Grammars, Generalized Phrase Structure Grammars, and Tree Adjoining Grammars. Those formalisms then started to put a stress on lexicon, and were called as lexicalist (or lexicalized) grammars. Representative examples of lexicalist grammars were Head-driven Phrase Structure Grammars (HPSG) and Lexicalized Tree Adjoining Grammars (LTAG). While grammars and lexicons were two major linguistic resources of syntactic processing of natural languages, lexicons began to play an important role in language processing.

Things have changed from early 90's, when large scale language resources became available and corpus-based research started to dominate almost all aspects of natural language processing (NLP). Part-of-speech taggers and syntactic parsers are the most well-studied topics in corpus-based research. Various parsers, based either on phrase structure grammars or on dependency structures, have been developed, applying various machine learning techniques on syntactically annotated corpora. State-of-the-art parsers developed in this way have achieved very good performance. Those trends are also beneficial to lexicalist grammars since parsing with those grammar formalisms is amenable to phrase structure-based parsing through abstraction of grammatical schemata or a derivation process with those grammar formalism (i.e., a derivation tree) can be considered to correspond to a word dependency tree.

Recent trends in NLP have started to target diversely spread areas that require semantic and pragmatic information. Some areas like social media analysis, such as twitter or blog text analysis, have a more preference to getting semantic or sentiment information than syntactic information. Though this trend is attracting people's attention and is getting growing importance, still syntactic analysis keeps to play an important role. Simple extension of annotated corpora and lexical statistics will not be able to skyrocket parsers' performance. Improvement of parsing accuracy especially that of long sentences requires to tackle problems that are not on the current main stream of parser development.

In this talk, I will take up three issues that lie between grammars and lexicons: Coordination structures, multiword expressions and complex sentence patterns. I will first give a brief overview of syntactic processing in past two/three decades, then will talk about the issues one by one especially about our experiences related with them. Finally, I will consider future directions of sentence analysis taking those into account.

Coordination Structures

Coordination Structures are well-known and notorious phenomena observed in all languages, and especially in long sentences. Not only pairs of phrases of the same category but also pairs of any sequences or words that are *similar* in some sense can be coordinated. No grammar formalisms, except for Categorial Grammars, can give a comprehensive account and appropriate representation for coordination structures.

There is a proposal to use dynamic programming matching to find coordination structures as they tend to consist of similar sequences of words or phrases. One problem, however, is: When they are coordinated, some constructions such as noun phrases or sequences of complements for a predicates usually have similar structures, other constructions such as verb phrases or compound sentences may have very different structures. Another problem is: A coordination structure may be embedded in another coordination structure while they cannot overlap each other.

I will give our experiences to handle embedded coordination structures and our experiments to see how coordination structure information helps improve parsing accuracy. Through those, I will talk about our findings.

Multiword Expressions

Multiword expressions (MWEs) are those consisting of multiple words that have non-compositional and/or idiosyncratic interpretations. Some of them, which appear in fixed forms, should be registered in a dictionary. However, there are other types of MWEs that have syntactic flexibilities. There are a series of workshops devoted to MWEs (http://multiword.sourceforge.net/).

Although construction of MWE lexicons and MWE annotated corpora is done in some languages such as French and Swedish, no large scale English MWE lexicon and MWE annotated corpus have been developed. Some of the MWEs have non-standard POS patterns and behave unpredictably from the constituent words, many of them should be registered in dictionaries for language processing.

I will give an overview of language analysis research with MWEs, and will give our current attempt to construct an English MWE dictionary and its application to Part-of-speech tagging.

Complex Sentence Patterns

Simple sentences in a language have a rather uniform construction. However, there are a variety of structures in complex sentences in any language. Subordinate structures and embedded clauses are typical structures of complex sentences, and those structures could be produced in a recursive manner, making an analysis of such structures very difficult. There are also some complex sentence patterns that are difficult to define in existing grammar formalisms. Such complex sentences are also very difficult to parse in existing parsing algorithms since they usually parse a sentence in a bottom-up manner assuming some type of locality.

I will talk about our recent experiments to find subordinate and embedded clause patterns in an auto-parsed English corpus. Although there are a huge number of complex sentence patterns, once they are attempted to merge into a smaller number of patterns by ignoring redundant phrases and punctuations we found that a small number of complex sentence patterns can have a very wide coverage of whole complex sentences. I will introduce the results of our experiments and will discuss further possibilities of extracting wider types of complex sentence patterns.

Considerations and Conclusions

The issues in sentence analysis discussed in this article are the remaining “things” we need to tackle between standard grammars and lexicons. The main difficulty related with these issues is that they are intermingling phenomena with the standard syntactic analysis. Knowing coordination structures, multiword expressions and complex sentence patters in advance in a given sentence is definitely useful to sentence parsing, while identifying those structures requires some syntactic analysis.

A natural conclusion is joint analysis of syntactic parsing and those specific constructions. There have been a number of proposals for joint processing of different levels of language processing, such as joint POS tagging and phrase/NE chunking, joint POS tagging and parsing, joint syntactic and semantic parsing, and so on. It is important and valuable to seek for methods of joint processing of syntax and the constructions taken up in this article.

Another important topic is how to acquire and represent the knowledge or expressions in a comprehensible and reusable format since those phenomena should be analyzed not only an independent manner but also in an integrated module in other language processing systems and tools. The know-how of extraction, construction and representation of those resources should be transferable over languages.

Acknowledgments

I would like to express my sincere appreciation to the staff and students in our laboratory for their cooperation and valuable discussions.