What Do Computational Linguists Need to Know about Linguistics?

Robert C. Moore
Microsoft Research
Redmond, Washington, USA
bobmoore@microsoft.com

Abstract

In this position paper, we argue that although the data-driven, empirical paradigm for computational linguistics seems to be the best way forward at the moment, a thorough grounding in descriptive linguistics is still needed to do competent work in the field. Examples are given of how knowledge of linguistic phenomena leads to understanding the limitations of particular statistical models and to better feature selection for such models.

Over the last twenty years, the field of computational linguistics has undergone a dramatic shift in focus from hand encoding linguistic facts in computer-oriented formalisms to applying statistical analysis and machine learning techniques to large linguistic corpora. Speaking as someone who has worked with both approaches, I believe that this change has been largely for the good, but I do not intend to argue that point here. Instead, I wish to consider what computational linguists (if it is still appropriate to call them that) need to know about linguistics, in order to work most productively within the current data-driven paradigm.

My view is that, while computational linguists may not need to know the details of particular linguistic theories (e.g., minimalism, LFG, HPSG), they do need to have an extensive understanding of the phenomena of language at a descriptive level. I can think of at least two somewhat distinct applications of this sort of knowledge in empirical computational linguistics.

One application is to understand the structural limitations of particular types of statistical models. For example, a descriptive generalization about language is that coordinated structures tend to be interpreted in such a way as to maximize structural parallelism. Thus, in the phrase “young men and women”, “young” would normally be interpreted as applying to both “men” and “women”, but in the phrase “young men and intelligent women”, “young” would normally be interpreted as applying only to “men”. Although both interpretations are structurally possible for both phrases, the preferred interpretations are the ones that maximize structural parallelism. This is a phenomenon that is not describable in a general way in a simple statistical model in the form of a probabilistic context-free grammar (PCFG). We could enumerate many specific cases by making fine-grained distinctions in the nonterminals of the grammar, but the tendency to favor parallel coordinated structures in general would not be expressed. This is not necessarily fatal to successful engineering applications of PCFGs, but a competent computational linguist should understand what the limitations of the formalism are.

Let me give another example from the notoriously empirical field of statistical machine translation (SMT). At least some linguistic structure has been creeping back into SMT recently in the form of hierarchical translation models, many of which can be viewed as instances of synchronous probabilistic (or more generally, weighted) context-free grammars (SPCFGs). This approach seems quite promising, but since it is based on a bilingual version of PCFGs, not only does it share the limitations of monolingual PCFGs alluded to above, but it also has additional structural limitations in the kind of generalizations over types of bilingual mappings it can model.

My favorite example of such a limitation is the translation of constituent (i.e., “WH”) questions between languages that move questioned constituents to the front of the question (“WH-movement”) and those that leave the questioned constituents *in situ*. English is an example of the former type of language, and Chinese (so I am told) is an example of the latter. If we wanted to make a model of question translation from Chi-

Proceedings of the EACL 2009 Workshop on the Interaction between Linguistics and Computational Linguistics, pages 41–42,
Athens, Greece, 30 March, 2009. ©2009 Association for Computational Linguistics

nese to English, we would like it to represent in a unitary (or at least finitary) way the generalization, "Translate the questioned constituent from Chinese to English and move it to the front of the English sentence being constructed." This generalization cannot be expressed in an SPCFG, because this type of model allows reordering to take place only among siblings of the same parent in the constituent structure. Fronting a questioned constituent, however, typically requires moving an embedded constituent up several levels in the constituent structure. While we can express specific instances of this type of movement using an SPCFG by flattening the intervening structure, we cannot hope to capture the generalization in full because WH-movement in English is famously unbounded, as in "What translation formalism did Moore claim to show that WH-movement could not be modeled in?"

In addition to providing a basis for understanding the limitations of what phenomena various statistical models can capture, a good knowledge of descriptive linguistics is also very useful as a source of features in statistical models. A good example of this comes from acoustic modeling in speech recognition. Acoustic models in speech recognition are typically composed of sequences of "phone" models, where a phone corresponds approximately to the linguistic unit of a phoneme. For good recognition performance, however, phone models need to be contextualized according to the other phones around them. Commonly, "triphone" models are used, in which a separate model is used for each combination of the phone preceding and following the phone being modeled. This can require over 100000 distinct models, depending on how many triphones are possible in a given language, which creates a sparse data problem for statistical estimation, since many of the possible combinations are only rarely observed.

One response to this sparse data problem is to cluster the states of the triphone models to reduce the number of separate models that need to be estimated, and an effective way to do this is to use decision trees. Using a decision tree clustering procedure, the set of all possible triphones is recursively split on relevant features of the triphone. At each decision point, the feature chosen for splitting is the one that produces the greatest improvement in the resulting model. But what features should be used in such a decision tree? I once heard a leading speech recognition engineer say that he chose his feature set by including all the features he could find in the linguistic phonetics literature. Given that feature set, the decision tree learning procedure decided which ones to actually use, and in what order.

The examples presented above illustrate some of the kinds of linguistic knowledge that a competent computational linguist needs to know in order perform research at the highest level. I am concerned that many of the students currently graduating in the field do not seem to have received sufficient exposure to the structure of language at this level of detail. For instance, a few years ago I pointed out the problem of modeling question translation between Chinese and English to one of the brightest young researchers working with SPCFGs, and the problem had never occurred to him, even though he was a fluent speaker of both languages. I am sure this would be one of the first things that would occur to anyone brought up on the debates of the 1980s about the limitations of context-free grammar, upon first exposure to the SPCFG formalism. So, although I am a firm believer that the data-driven empirical approach computational linguistics will remain the most fruitful research paradigm for the foreseeable future, I also think that researchers need a firm grounding in descriptive linguistics.