Structure in Linguistics

Linas Vepstas
Novamente, OpenCog Project

1. Introduction

A combination of increasing computer power and commercial utility will attract new attention to the field of corpus linguistics, and lead an invasion of computational linguistics principles. Access to computer resources allows large data-driven experiments to be performed using large collections of linguistic data, while commercial utility introduces a pragmatic engineering ethos where workable methods and practical results are more important than theoretical debates. The corpus provides the fertile ground on which this growth of interest can take place.

Half a century of computational linguistics has revealed that building AI/NLP systems by hand is a difficult and seemingly insurmountable task. However, the process of hand-building seems understood well enough that parts of it can be automated — that is, grammars can be learned by machine-learning systems, and semantic content can be automatically discerned. The availability of a large corpus appears to be crucial for this task. In this sense, the distinction of corpus linguistics and computational linguistics is now blurred (see Gries, this volume; Wilks, this volume).

The use of experimental techniques to discover new linguistic facts hidden in large corpora have provided many new discoveries. However, some (Varela et al. 1991, Pfeifer & Scheier 1999) suggest that there are even stronger ways to discover linguistic and semantic structure: "linguistics by embodiment". For humans, it is easier to learn something by doing it, as opposed to just reading about it. Similarly, the embodiment hypothesis is that a software agent can more easily learn semantic communication by having a virtual, 3D body allowing linguistic behaviors to be correlated with visual and motor senses. This raises important theoretical and philosophical questions about "concepts" and "meaning": Is meaning confined to the evidence attainable from a corpus (Teubert, this volume), or is there a sublime aspect of meaning that is accessible only by doing, feeling, sensing? In humans, internal feelings and sensations are not directly accessible to outside observers, and so such debates can be rendered moot by lack of evidence. The conceit of

International Journal of Corpus Linguistics 15:2 (2010), **357–362**. DOI 10.1075/ijcl.15.2.06vep
ISSN 1384–6655 / E-ISSN 1569–9811

corpus linguistics, that meaning is negotiated between speakers (Williams, this volume; Teubert, this volume), and is located only in the discourse (Teubert, this volume), has reigned unchallenged. But what happens when one has access to a robotic linguistic actor, whose internal states can be probed and examined, where the sublime and ineffable can be viewed and measured? This possibility even turns corpus linguistics on its head: does it make sense to collect and analyze a corpus of humans talking to automated bankers and pizza ordering systems? Surely, the fact that the software programming for the pizza ordering system is available seems to obviate the need to sample its utterances. And yet, a sampling of the confusions and digressions that occur during chat between humans and robotic systems can be commercially valuable: the efficiency and accuracy of these systems can be improved by studying how communication breaks down.

The goal of the rest of this screed is to sketch not only how linguistic analysis is performed automatically, but how deeply layered and abstract it has become. Two points emerge: collocation and corpus are key elements, and forward progress requires clever observation of subtle linguistic phenomena by human linguists.

2. Collocation from a mathematical point of view

The starting point for automated linguistic learning is collocation. One of the simplest analyses one can perform is to discover important bigrams by means of mutual information (Yuret 1998). While a simple count of probabilities might show which word pairs (or N-grams) occur frequently, mutual information gives a different view, as it uncovers idiomatic expressions and "set phrases": collections of words that are used together in a relatively invariant fashion. Examples of high mutual information word pairs are *Johns Hopkins* and *jet propulsion*, both scoring an MI of about 15 in one dataset (Vepstas 2008). The most striking result from looking over lists of high-mutual-information word pairs is that most are the names of conceptual entities.

Mutual information between word pairs provides the starting point for a class of parsers such as the Minimum Spanning Tree Parser (McDonald & Pereira 2005). These parsers typically are dependency parsers that discover dependency relations using only "unsupervised" training. The key word here is "unsupervised": rather than providing the learning algorithm with linguist-annotated datasets, the parser "learns" merely by observing text "au naturel", without annotations. This is in contrast to many other parsers built on statistical techniques, which require an annotated corpus as input. By eschewing annotation, it stays true to one of the core ideals of corpus linguistics: "trust the text".

Sometimes it is imagined that statistical techniques stop here: anyone who has played with (or read about) Markov chains built from N-grams knows that such systems are adept at generating quasi-grammatical nonsense text. But this does not imply that machine learning techniques cannot advance further: it merely means that a single level of analysis has only limited power, and must be aided by further layers providing additional structure and constraints. So, for example, although a naive Markov chain might allow nonsense word combinations, a higher order structure discriminator or "layer" would note that certain frequently-generated patterns are rarely or never observed in an actual corpus, and prohibit their generation (alternatively, balk at creating a parse when presented with such ungrammatical input). Examples of such "layering" are presented below.

Another important point is that the observation of structures in corpora can be used to validate "mentalistic" theories of grammar. The whole point of traditional, hand-built parsers was to demonstrate that one could model certain high-frequency word occurrences with a relatively small set of rules — that these rules can differentiate between grammatical and ungrammatical text. The "small" set of rules capture and distil the contents of a much larger corpus. Thus, for example, high mutual information word pairs help validate the "mentalistic" notion of a "dependency" parser. But this is a two-way street: theories based on practical experience can serve as a guide for the kinds of structures and patterns that one might be able to automatically mine, using unsupervised techniques, from a text corpus. Thus, for example, the "Meaning-Text Theory" developed by Igor Mel'cuk and others (Mel'cuk & Polguere 1987, Steele 1990) provides a description of 'set phrases', deep semantic structures, and of the interesting notion of 'lexical functions'. Perhaps one might be able to discover these by automated means, and conversely, use these automatically discovered structures to perform more accurate parsing and even analysis of semantic content.

Such attempts at automated, unsupervised learning of deeper structure are already being done, albeit without any theoretical baggage in tow. Prominent examples include Dekang Lin's work on the automatic discovery of synonymous phrases (Lin & Pantel 2001), and Poon & Domingos' (2009) extension of these ideas to the automated discovery of synonymous relations of higher N-arity. The earlier work by Lin applies a fairly straightforward statistical analysis to discover co-occurring phrases and thus deduce their synonymy.

However, it is important to note that Lin's analysis is "straightforward" only because a certain amount of heavy lifting was already performed by parsing the input corpus. That is, rather than applying statistical techniques on raw N-grams, the text is first analyzed with a dependency parser; it is the output of the parser that is subject to statistical analysis. Crudely speaking, "collocations of parses" are found.

This provides an example of the kind of "layered" approach that seems to be necessary to tease out the deeper structures in text — it is not possible to define a single, simple statistical measure that will magically reveal hidden structure in text. Indeed, the complexity of techniques can quickly escalate: although the work of Poon & Domingos (2009) can be seen as a kind of generalization of that of Lin, it employs a considerably more dense and intellectually challenging mathematical framework to do so. Nonetheless, it takes as its core the idea of "trust the text" — neither linguistics nor meaning are "engineered into the system" — there are no ad-hoc English-language-specific tweaks in the code, and, indeed, almost no linguistic theory whatsoever, beyond the core requirement that dependency parsing provide the input.

And so we've painted a picture of a layered cake of analysis, starting with a corpus, passing through mutual information to a dependency parser, from which further relationships are teased out using progressively more abstract principles. There are certainly other recipes for that cake. Phrase-structure-driven parsers can also be learned by statistical techniques, although the *unsupervised* discovery of such grammars is more mathematically challenging (i.e. requiring a deeper understanding of topics in computer science). That "large scale knowledge extraction" is possible from such grammars is demonstrated by Michael & Valiant (2008).

A major blot on the above argument is that the nature of the driving force behind the results is the urge to discover some practical, usable, functional technique for automatically discovering and manipulating "knowledge". This is a laudable goal, for doing so will presumably shed light on the nature of "meaning" and "concepts", as well as be of considerable economic benefit. Yet, in the mad rush towards automated knowledge extraction and manipulation, finer and more subtle linguistic phenomena are left on the wayside. The overall attack is still brute-force: the knowledge to be extracted is often no deeper than an analysis of the sentence *Aristotle is a man*, and often requires little more linguistic analysis than that needed to parse that sentence. This is partly because the methods are still crude, but is also partly cultural: engineers and mathematicians usually have little appreciation of linguistic tradition arising from the Humanities. The genres frequently analyzed are financial or biomedical texts; fine literature and poetry are roundly ignored. I can only hope that perhaps some of these techniques can be refocused on more subtle phenomena, under the guidance or leadership from scholars in the more humanistic side of linguistics.

The trend towards ever-more complex analysis seems inexorable. But this does not mean that every simple, fast, easy way of analyzing text has been discovered. Recent work from Google and NEC labs titled "Polynomial Semantic Indexing" (Bing et al. 2009) demonstrates how a fairly simple, shallow mathematical technique, which assumes no syntactic structure in the text whatsoever, can nonetheless discover considerable amounts of semantic content. While the technique is

hardly any deeper than that of computing co-occurrence probabilities or mutual information, it is notable in that it requires a miniscule amount of CPU time, as compared to other "obvious" approaches, or even the act of parsing text.

3. Conclusion

In the connectionist view of semantics, "meaning" exists only in patterns and relations and connections: meaning is expressed in linguistic dialog. In the case of embodied actors, meaning can be expressed in action; but again, the action is in reference to some thing. There is no atomic kernel to meaning, there is no "there", there. Rather, meaning is tied into the expression and articulation of structure.

But what is meant by "structure" and "connections"? There are multiple candidates for things that could be "structure" in linguistics. One obvious candidate is "lexis" — that grand lexicon in the sky, where all words, phrases and idioms are defined maximally in relation to one another. But there is also the structure commonly known as "syntax" — a set of rules governing the positional ordering of words. Cognitive linguistics posits that there are even deep structures of various sorts (such as the "lexical functions" of Meaning-Text Theory). It would seem that each of these different kinds of structures can be discerned to some degree or another using automatic, unsupervised machine-learning techniques, taking only a bare, naked corpus of text as input.

To put it crudely, collocation need not always be done "by hand": let the computer do it, and look at what it does. Don't just look at what words are next to each other, but look at what structures are next to each other. Future effort at the intersection of computational and corpus linguistics is in the discovery of more subtle structure and the invention of new techniques to expose and manipulate it. Insofar as current structures only capture an approximation of human discourse, then clearly more work remains to be done.

Perhaps to the reader, this layering and interaction of different algorithms to explain linguistic phenomena is reminiscent of the use of epicycles on epicycles to correct for circular planetary orbits. Perhaps it is. But at this stage, it is still too early to tell. Certainly, the parsers of a few decades back were perhaps less than impressive. Language has a lot of structure, and capturing that structure with hand-coded rules proved to be an overwhelming task. Yet, the structure is there, unclear as it may often be. What we have now, that we didn't have back then, are the tools to raise the exploration to a "meta" level: the tools themselves find rules and structure in the corpus — and now we can try to understand what sort of rules and structure they are finding, and why some kinds of structure are invisible to some kinds of tools, and what it is that makes one tool better than another.

References

Bing, B. et al. 2009: online. "Polynomial semantic indexing". In *Proceedings of the 2009 Conference in Advances in Neural Information Processing Systems 22*. Available at: http://books.nips.cc/papers/files/nips22/NIPS2009_0881.pdf (accessed April 2010).

Lin, D. & Pantel, P. 2001. "Discovery of inference rules from text". In *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'01)*. USA: ACM Press, 323–328.

McDonald, R. & Pereira, F. 2005: online. "Minimum-Spanning Tree Parser". Available at: http://www.seas.upenn.edu/~strctlrn/MSTParser/MSTParser.html (accessed April 2010).

Mel'čuk, I. A. & Polguère, A. 1987. "A formal lexicon in the meaning-text theory: (Or how to do lexica with words)". *Computational Linguistics*, 13 (3–4), 261–275.

Michael, L. & Valiant, L. G. 2008. "A first experimental demonstration of massive knowledge infusion". In *Proceedings of the 11th International Conference on Principles of Knowledge Representation and Reasoning, Sydney, Australia, 16–20 September 2008*, 378–389.

Pfeifer, R. & Scheier, C. 1999. *Understanding Intelligence*. Cambridge, MA: The MIT Press.

Poon, H. & Domingos, P. 2009 "Unsupervised semantic parsing". In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*. Singapore: Association for Computational Linguistics, 1–10. Also available at: http://www.aclweb.org/anthology/D/D09/D09–1001 (accessed April 2010).

Steele, J. (Ed.) 1990. *Meaning-Text Theory: Linguistics, Lexicography, and Implications*. Canada: University of Ottawa Press.

Varela, F. J., Thompson, E. & Rosch, E. 1991. *The Embodied Mind: Cognitive Science and Human Experience*. Cambridge, MA: The MIT Press.

Vepstas, L. 2008: online. "Linas' collection of NLP data". Available at: http://gnucash.org/linas/nlp/ (accessed April 2010).

Yuret, D. 1998. *Discovery of Linguistic Relations Using Lexical Attraction*. PhD Thesis, Massachusetts Institute of Technology. Also available at: http://arxiv.org/PS_cache/cmp-lg/pdf/9805/9805009v1.pdf (accessed April 2010).

Author's address

Linas Vepstas
Novamente, OpenCog Project
1518 Enfield Road
Austin TX 78703

linasvepstas@gmail.com

About the author

Dr. Linas Vepstas is a relative newcomer to the field of linguistics, and thus expresses a naive view of the field, likely to be grating to old hands. Worse, his doctorate is in physics and mathematics, and so is doubly cursed by those who have come to linguistics as a discipline of the humanities. He hopes to make up for these sins with enthusiasm, striving for some state of grace.