Can Statistical Parsers WOW! You: A Cognitive Assessment

Sandiway Fong (sandiway@email.arizona.edu)
Departments of Linguistics and Computer Science, University of Arizona, Douglass 200E
Tucson, AZ 85721 USA

Robert C. Berwick (berwick@csail.mit.edu)
Departments of EECS and Brain and Cognitive Sciences, MIT, 32D-728, 77 Massachusetts Ave.
Cambridge, MA 02139 USA

One of the presumptive goals of cognitive science is a faithful implementation of linguistic knowledge. In the case of statistically-based parsers trained on large corpora, such as the Penn Tree Bank Wall Street Journal sentences (PTB), we might pose this goal as follows: How closely do statistical parsers trained on this data replicate human acquisition and knowledge of language rather than the standardly-used precision/recall information retrieval engineering metrics — a kind of cognitive "Turing test", in the domain of parsing. One would expect a cognitively faithful model to to be able to more easily acquire *natural* (i.e., human) grammatical rules than grammatical patterns not attested in any natural language.

One well-known such constraint is that no natural language makes use of "counting," in the sense of basing grammatical rules on a particular number of tokens rather than on syntactic structure. For example, there is apparently no natural language in which the negation of a declarative sentence is formed by inserting a particular morpheme as the fourth word from the beginning of the corresponding declarative (as opposed to the sixth or seventh word). Such examples have been examined previously, including (1) a comparison of artificial language learning by adults and the linguistic 'savant' Christopher individual, e.g., (Smith, Tsimpl, & Ouhalla, 1993), where experimenters constructed an artificial language Epun in which a particular emphatic form was based on counting rather than syntactic structure; and (2) an fMRI study of the acquisition of natural and artificial languages by adults (Musso & Moro, 2003), who constructed artificial examples with negation of exactly the sort described above, with the fourth word marking negation (e.g., *Paolo mangia la no pera*–"Paolo eats the no pear").

Both cited experiments reported differences in the ease of learning "unnatural" vs. "natural" grammatical rules: in the first experiment, the "unnatural" rules were essentially impossible to learn by the compromised but otherwise grammatically adept individual and extremely difficult for normal adults. In the second experiment, quite different fMRI regions were activated while learning "unnatural" rules, in contrast to the regions involved in learning natural rules (Broca's area in the latter case). Both results, along with others of a similar nature, have been interpreted to indicate that the learning of "unnatural" counting constructions invokes "problem-solving" cognitive machinery distinct from that of natural language.

The key question we address in this paper is whether statistically-trained parsers mirror these paradigmatic results about human language competence: do they also find that learning "unnatural" counting constructions is extremely difficult, or whether, in contrast to humans, they easily learn such rules. To test this claim, we constructed from the 5507 passive examples in the PTB an artificial set of training data sentences in which English passive sentences were transformed into examples as close those in the Smith and in the Musso experiments as possible. We removed the passive morphology that ordinarily indicates an English passive (e.g., *was named*) and replaced it with a special marker word **Wow!** at the fourth position from the start of the sentence (S). We then re-trained the Bikel implementation of the Collins statistical parser on this new data,(Collins, 2003),(Bikel, 2004) and tested its performance on the standard held-out test set (PTB section 23) of 327 similarly transformed passive test sentences, comparing the parser's performance on the same training/test umodified passive dataset.

Results. We found that the parser had a somewhat greater difficulty in learning **Wow!**-transformed examples as compared to normal passives, a precision + recall (F-measure) score of 85.85% vs. 88.49%. However, this still indicates a very high level of correct learning, despite the drop: such a number is typically taken as indicative of high success in statistical learning(Manning, 2003). The ability of the statistical system to learn a counting-type rule this well thus contrasts sharply with human performance. Further, given the fact that such parsers are inherently phrase-based, it is actually quite difficult to achieve perfect fourth-word insertion, which introduces "noise" into the training data. Thus part of the drop might be attributable to this factor. We conclude therefore, tentatively, that at least this statistically-based parser does not perform in a manner comparable to humans: it can learn "counting" rules with an accuracy approaching that of a "natural" rules. In this respect, such statistical parsers are *not* cognitively faithful.

Comments. Since all such parsers are constituent-based, extremely accurate generalization is probably not achievable for this grammatical construction. Perfect 4th word insertion would mean acquiring a very large number of different patterns containing **Wow!** – since the 4th word could actually appear at almost any point in the derivation. For example, it is simply not possible to insert **Wow!** as the 4th word of every clause consistently with respect to syntactic

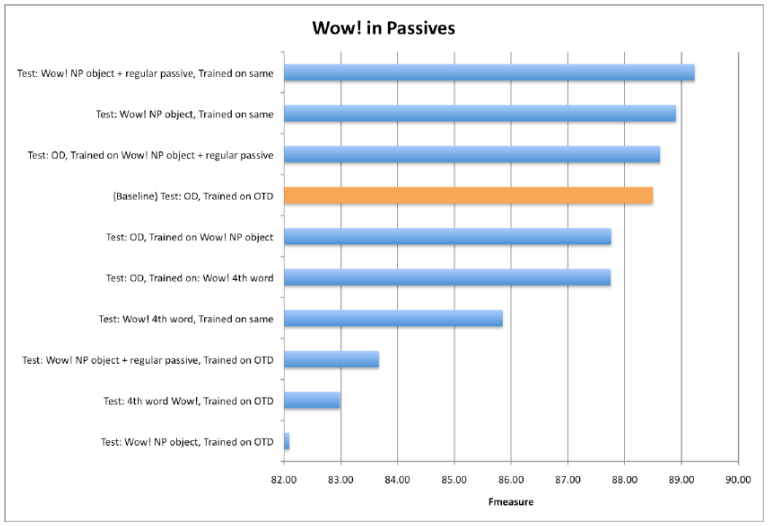

Figure 1: Results for **Wow!** experiments. Combined precision+recall (F-measure) percent parsing accuracy on section 23 PTB test data is displayed on the X-axis. The Y-axis displays the complete set of experimental manipulations performed, of which just two are the focus here. The original "gold standard" parsing test results are indicated by the gold horizontal bar notated on the Y-axis as "Baseline," at 88.6% accuracy. The direct **Wow!** fourth-word insertion is given just by the third blue bar below the gold bar, notated "Test: Wow! 4th word. Trained on same" with an F-measure score of 85.85% accuracy.

level, because: (i) there may not be $n-1$, i.e. 3, pre-existing words in a clause: John was arrested/John arrests; (ii) unless we can represent discontinous constituents, **Wow!** cannot be consistently inserted as exactly the n^{th} word and be present at the same syntactic level. For example, given the original sentence, *Although preliminary findings were reported more than a year ago , the latest results appear in today 's New England Journal of Medicine* ., note that if we insert **Wow!** strictly as the fourth word **at the same syntactic level** as the S, since *although* appears as an adverbial one clause up, **Wow!** would be inserted *after* the verb *reported*. Such irregularities might the cause of the somewhat decreased parsing accuracy that we observed. Finally, in other experiments (see blue bars above gold bar in the figure), we found, not surprisingly, that inserting **Wow!** to explicitly mark the object Noun Phrase "trace" that is linked to the Subject NP could substantially improve baseline performance, since this "regularizes" predicate-argument structure, rendering it less noisy for statistical learning.

Details and Data. The training set consisted of 39,832 examples from the PTB (Wall Street Journal, sections 02–22), with 5507 passive examples.

The test set consisted of passive sentences only from section 23 of the PTB (327 sentences out of a total of 2416 sentences, i.e., 14% of the total number of sentences in section 23). In the figure below, OTD denotes the original training data, i.e. sections 2 through 21 prior to removal of passive morphology and insertion of **Wow!** as the 4th word and/or other manipulation. (Results not marked with OTD are obtained with training data comprising sections 2 through 21 of the PTB with passive morphology removed but with **Wow!** inserted as the 4th word in the S, signaling the presence of the passive construction.) An example of an original sentence in the PTB followed by its transformed counterpart follows:

```
Rudolph Agnew , 55 years old and former chairman of
Consolidated Gold Fields PLC , was named a nonexecutive
director of this British industrial conglomerate .
```

The transformed new sentence with **Wow!** inserted as fourth word from the left is as follows:

```
Rudolph Agnew , Wow! 55 years old and former chairman
of Consolidated Gold Fields PLC , named a nonexecutive
director of this British industrial conglomerate .
```

Acknowledgments

We thank Michael Coen, Igor Malioutov, Robert Speer, and Beracah Yankama for assistance and valuable suggestions.

References

Bikel, D. M. (2004). Intricacies of Collins' parsing model. *Computational Linguistics*, *30*(4), 479–511.

Collins, M. (2003). Head-driven statistical models for natural language parsing. *Comput. Linguist.*, *29*(4), 589–637.

Manning, C. (2003). Probabilistic syntax. In R. Bod, J. Hay, & S. Jannedy (Eds.), *Probabilistic linguistics* (pp. 289–341). Cambridge, Massachusetts: MIT Press.

Musso, & Moro, A. (2003). Broca's area and the language instinct. *Nature Neuroscience*, *6*(7), 774–781.

Smith, N., Tsimpl, I.-M., & Ouhalla, J. (1993). Learning the impossible: The acquisition of possible and impossible languages by a polyglot savant. *Lingua*, *91*, 279–347.