

# 5 Questions and answers about prospects for statistical NLP

Chris Brew and Mike White





Q: Can you beat  
Google?



# Q: Can you beat Google?

- All this fuss about parsing is a waste of time, given that the rising tide of unanalysed words improves Google search far more than anything else.



# Q: Can you beat Google?

- All this fuss about parsing is a waste of time, given that the rising tide of unanalysed words improves Google search far more than anything else.
- This is true, but the plus is in their ability to use very large-scale data, not in the lack of analysis. So no, we can't beat them by doing something different from what they do, but we can use language insight to generate other features they can put into their data mills.



# Q: Can you beat Google?

- All this fuss about parsing is a waste of time, given that the rising tide of unanalysed words improves Google search far more than anything else.
- This is true, but the plus is in their ability to use very large-scale data, not in the lack of analysis. So no, we can't beat them by doing something different from what they do, but we can use language insight to generate other features they can put into their data mills.
- Corollary: gains from applying linguistic ideas will only accrue if we make sure that in using them we don't abandon the strengths of the simpler.



**Q: Isn't parsing just too slow?**



Q: Isn't parsing just too slow?

- Powerset was offering to use parsing to search the web, but their demo is just Wikipedia.



# Q: Isn't parsing just too slow?

- Powerset was offering to use parsing to search the web, but their demo is just Wikipedia.
- OK, but Clark and Curran's parser is fast enough that we are able to parse the English Gigaword (1000 million words) in a day or two. (Credit: OSC's big cluster of Linux machines.)



**Q: Must I work on newspapers?**



Q: Must I work on newspapers?

- I hate the Wall Street Journal, please can I work on something fun instead?



# Q: Must I work on newspapers?

- I hate the Wall Street Journal, please can I work on something fun instead?
- Sorry, you do have to work on the Wall Street Journal if you want to play in this area, but c.f. for example work in Manchester and Tokyo on domain dependence and domain adaptation.



**Q: How does this  
involve Linguistics?**



# Q: How does this involve Linguistics?

- I don't believe you can get insight out of all this data.



# Q: How does this involve Linguistics?

- I don't believe you can get insight out of all this data.
- Fair enough. Onus is on us. My hope is that some meaningful approximation to (part of) Logical Forms can be obtained from large data.



# Q: How does this involve Linguistics?

- I don't believe you can get insight out of all this data.
- Fair enough. Onus is on us. My hope is that some meaningful approximation to (part of) Logical Forms can be obtained from large data.
- Jianguo Li, Kirk Baker and I are working on ways to use parse outputs to understand/extend Beth Levin's claims about semantics of verb classes.



Q: What's the next step?



# Q: What's the next step?

- Existing successful parsers (e.g. Collins) include an implicit model of verb subcategorization. Why bother with anything else?



# Q: What's the next step?

- Existing successful parsers (e.g. Collins) include an implicit model of verb subcategorization. Why bother with anything else?
- For different domains and unknown words, the linguist-derived organization of verbs into semantically motivated classes offers hints of the kind of generalizations that the data might support. Can we use parse outputs to get evidence out of really big datasets? Can we apply cloud computing ideas to get this on really really big datasets?