Speak My Language

Natural Language Processing With Ruby
What I’ll Cover

• Who I am

• Overview of some Ruby NLG and NLP libraries

• Emphasis on a few I’ve written: Linguistics, Ruby-WordNet, Ruby-LinkParser

• Lots of samples, Zenspider-style
What I Won’t Cover

• Hardcore computational linguistics
• DSLs
• Details about the algorithms behind the libraries
Who am I?
true
just another ruby hacker
• Rubyist since late 2000

• Author of some random stuff: Arrow, IO::Reactor, BlueCloth, CrossCase, Linguistics, Ruby-WordNet, MUES, Ruby-MemCache, Ruby-ODE, OOParser, Test::Unit::Mock, and a perpetual work-in-progress online game called FaerieMUD
• After 5 years, I finally get paid to hack Ruby all day!
Natural Language Processing
Using Ruby in Your English
Stemmable

- Implementation of the Porter Stemming Algorithm
- Author: Ray Pereda with improvements by Dave Thomas
Alia pts/ttyp1 [~/rubyconf2006]
51/542 {0}$
rubyreuters

- Text-categorizer using Reuters corpus
- Author: Mark Watson
Alia pts/ttyp1 [~/rubyconf2006]
78/568 {0}$
Oh well. At least it worked once.

After multiple takes.

Good thing I’m using Ryan’s magical presentation shell.
Chronic
(no, not *that* kind of “chronic”)

- A natural language date/time parser written in pure Ruby.
- Author: Tom Preston-Werner
irb(main):001:0>
Ruby-WordNet

- a Ruby interface to WordNet®.
- WordNet® is a lexical reference system invented at Princeton
- Author: me, ported from a similar Perl library by Dan Brian
• English nouns, verbs, adjectives and adverbs are organized into synonym sets, each representing one underlying lexical concept.

• Different relations link the synonym sets.

  • Hypernyms (superclass), hyponyms (subclasses), Meronyms (composed-of), Holonyms (composing), Antonyms (opposites), and many more
irb(main):001:0>
Ruby-LinkParser

- A Ruby binding to the Link Grammar library
- Link Grammar was written at Carnegie Mellon by Daniel Sleator, Davy Temperley, John Lafferty
• I wrote what I think was the original port in C

• A friend of mine (Martin Chase) attempted to write one in pure Ruby

• Newer version (again in C) takes advantage of a cleaned-up version released by the Abiword project
• There are now several other bindings, but I haven’t tried any of them
irb(main):001:0>
Other Stuff

- A bunch of nifty little NLP tools by UTIYAMA Masao

http://www2.nict.go.jp/x/x161/members/mutiyma/software.html
Natural Language Generation

Using English in Your Ruby
Rails Inflections

- Part of ActiveSupport
- Author: those Rails guys
- It’s what lets you say #pluralize and #singularize
• It’s good stuff!

• Very useful, saves awkward code

• Covers most common cases

• Small memory footprint

• Easily extensible/expandable
Rails Inflections

- ...but it does fall down sometimes
Ruby Linguistics

- A framework for building linguistic utilities for Ruby objects in any language.
- Author: me, on the shoulders of giants
• Heavier, more pendantic plural inflection

• Indefinite articles

• Numbers to words
• Infinitives

• Conjunctions
irb(main):001> require 'linguistics'; Linguistics.use( :en, :installProxy => true )
  => [String, Numeric, Array]

irb(main):002>
• Integrated WordNet and LinkParser
• Some things you can use this stuff for:
  • Command parsers
  • Improved search intelligence
  • Document categorization, auto-tagging
  • Fun
Alia pts/tty1 [~/rubyconf2006]
125/598 {0}$
Conclusion

• Get some of it at: http://deveiate.org/
• Other stuff via Rubyforge, RAA, etc.
• Thanks to: Damian Conway, Dan Brian, and Matz!