Updates from the ERG and Redwoods Treebanks

Dan Flickinger

Center for the Study of Language and Information

Stanford University

danf@stanford.edu

DELPH-IN Summit, Barcelona, 21 July 2009

Grammar Release Cycle

- Improve grammar
- Parse regression profiles, update treebanks, and correct errors
- Validate MRSs via *utool* for 'csli', 'mrs', and 'hike'
- Generate from 'csli', 'mrs', and 'hike'
- Parse corpora, update treebanks, and correct errors
- Rebuild parse ranking models (LOGON, WeScience)
- Update SEM-I (semantic interface)
- Verify updates for LexTypeDB
- Announce release

Recent additions to the ERG

- Chart mapping for preprocessing

  Numbers, measures, dates, etc.

- Improved unknown word handling for parsing and generation

- Expanded syntactic coverage

  Prepositional passives

  – *This is* **referred to** *as blackmail.*

  VP-modifying relatives

  – *Kim left, which bothered us.*

  Interleaving of VP modifiers and complements

  – *Kim discovered* **on that day** *a very interesting fact*

  Small clauses as sentence modifiers

  – *He bought the car, his brother objecting vigorously.*

Barcelona release of the ERG

- 208 syntactic rules
  - 39 coordination, 23 fragment, 22 head-adjunct, ...
- 67 lexical rules
  - 15 inflectional, 30 derivational, 22 punctuation
- 37,000 lexical stems (incl. 8600 proper names)
- 4200 types, incl. 863 lexical types
- 53,000 lines of TDL code, and 12,000 of comments

Treebanks for Regression Testing

Corpus type	Number of sentences	Avg. item length	Raw coverage	Treebanked coverage
CSLI/HP Test suite	1348	6.5	95.2%	92.5%
MRS	107	4.5	100%	100%
LOGON:hike	330	12.9	100%	99.4%
TREC	693	6.9	98.3%	97.5%
FraCaS	640	7.6	97.7%	97.0%

Treebanked corpora

Corpus type	Number of sentences	Avg. item length	Raw coverage	Treebanked coverage
Meeting/hotel scheduling	11660	7.5	96.8%	93.8%
E-commerce	5392	8.0	96.1%	93.0%
Dictionary defs.	10000	6.0	81.2%	75.5%
Norwegian tourism	10299	15.0	94.6%	87.9%
SemCor (part)	2501	18.0	91.8%	82.0%
Technical manuals	4000	12.5	86.8%	61.9%
Online user forum	578	12.5	85.5%	77.5%
Online essay	769	21.6	84.4%	65.1%
Chemistry papers	637	27.0	87.8%	65.3%
Wikipedia (Comp.Ling.)	8098	18.0	88.4%	77.3%

Corpus examples

- Meeting/hotel scheduling: VerbMobil
  Looks like we, need to schedule another meeting, in the next couple of weeks
- E-commerce: YY Software
  Don't ship the order and send me a refund immediately.
- Dictionary definitions: GCIDE
  Form: to shape, mold, or fashion into a certain state or condition;
- Norwegian tourism: LOGON
  If you would rather go fishing, there are opportunities in both Øvre Sjodalsvatn and Bessvatn.
- SemCor: Brown +
  Anyone's identification with an international struggle, whether warlike or peaceful, requires absurd oversimplification and intense emotional involvement.

- Online user forum: ILIAD
  Not sure if you ever got Linux installed dbessell, but this brings up a good point.
- Wikipedia: Computational Linguistics
  "'Computational linguistics"' is an [[interdisciplinary]] field dealing with the [[Statistics|statistical]] and/or rule-based modeling of [[natural language]] from a computational perspective.
- Chemistry papers: SciBorg
  By taking advantage of the growth steering properties of the OSCAR-COMPOUND film we were able to prepare nearly perfectly ordered hexagonal arrays of OSCARCOMPOUND clusters with a uniform distance of 4.5 nm between the particles.
- Technical manuals: CheckPoint
  Park tractor on flat level surface, shut engine off and place transmission in park.
- Online essay: "The Cathedral and the Bazaar"
  One key to understanding is to realize exactly why it is that the kind of bug report non-source-aware users normally turn in tends not to be very useful.

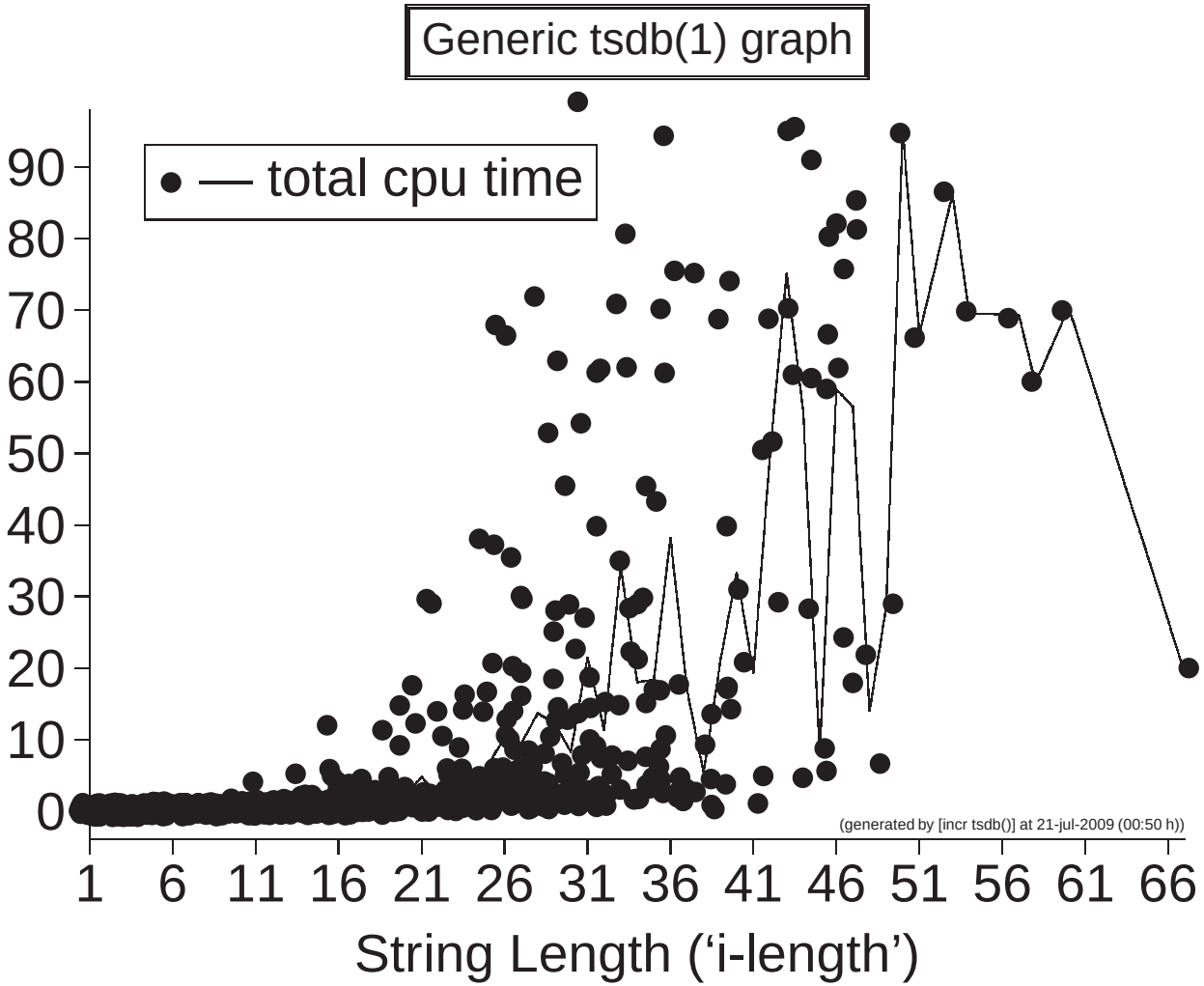
Parser and Grammar Configurations

- WeScience
  - PET: `-t -tsdb -yy -chart-mapping -packing -default-les=all -memlimit=1024 -sm=wescience.mem -timeout=100`
  - Preprocessor: `lkb::repp-for-pet ()`
  - POS tagger: TnT trained on WSJ
  - Resource limits: 100,000 edges, 500 parses per item
  - Roots: *root_strict root_informal root_frag root_inffrag*
- LOGON
  - PET: `-t -tsdb -yy -chart-mapping -packing -default-les=all -memlimit=1024 -sm=jhpstg.mem -timeout=60`
  - Preprocessor: `lkb::repp-for-pet ()`
  - No tagger
  - Resource limits: 100,000 edges, 500 parses per item
  - Roots: *root_strict root_informal root_frag root_inffrag*

'wescience/ws09/09-07-05/0907/gold' Performance Profile							
Aggregate	**items** $\sharp$	**etasks** ϕ	**filter** %	**edges** ϕ	**first** ϕ (s)	**total** ϕ (s)	**space** ϕ (kb)
$65 \leq$ *i-length* < 70	1	259370	98.2	31562	16.96	20.04	1050947
$60 \leq$ *i-length* < 65	1	662507	99.1	68440	0.00	70.40	1587859
$55 \leq$ *i-length* < 60	2	715278	98.1	79736	0.00	64.58	1511923
$50 \leq$ *i-length* < 55	4	660179	98.5	77418	0.00	79.39	1253729
$45 \leq$ *i-length* < 50	15	493976	98.3	58880	18.68	48.91	762035
$40 \leq$ *i-length* < 45	19	398528	98.3	50633	20.36	48.35	951913
$35 \leq$ *i-length* < 40	33	227253	98.3	29858	12.01	23.67	555502
$30 \leq$ *i-length* < 35	55	171916	98.3	25889	9.14	20.34	476570
$25 \leq$ *i-length* < 30	98	111852	98.2	17618	6.08	12.38	323402
$20 \leq$ *i-length* < 25	104	38262	98.2	8298	2.49	3.71	160938
$15 \leq$ *i-length* < 20	162	19504	98.2	5125	0.95	1.61	108095
$10 \leq$ *i-length* < 15	142	7159	98.1	2192	0.31	0.54	61700
$5 \leq$ *i-length* < 10	122	1496	98.2	583	0.08	0.12	30161
$0 \leq$ *i-length* < 5	180	144	98.1	76	0.01	0.02	11040
Total	**938**	**59973**	**98.3**	**9211**	**2.46**	**6.44**	**176021**

(generated by [incr tsdb()] at 21-jul-2009 (00:49 h))

Performance on WeScience 'ws09'



Next Steps

- Complete updates of Redwoods treebanks (5 days)
- Complete WeScience treebank (5000 remaining items, 5 weeks)
  Then recompute model, update treebank, and redo model
- Evaluate utility of ERG for grammar checking in education
  Sentences produced by children studying English
- Normalize rule names
  Propagate in treebanks and models