University of Washington (& beyond) site update

DELPH-IN Summit
July 29, 2013
St. Wendel, Germany

Emily M. Bender, Michael Wayne Goodman, Sanghoun Song, David Wax, Joshua Crowgey, Woodley Packard, Glenn Slayden, Zina Pozen, Megan Schneider, Prescott Klassen, Antske Fokkens, Ned Letcher

Previews

- Building a DELPH-IN Grammar for Lushootseed (Crowgey)
- MMT with ICONS and ACE (Song)
- Statistical Transfer (Goodman)
- Annotated Corpus for Phenomenon Construction (Letcher)
- CLIMB Update (Fokkens)
- Recap of UW Seminar on MRS and Other Semantic Representations (Bender)
- Boot-strapping Thai Parse Selection via ERG Translation Pairing and SVD-mapped Semantics (Slayden)
- Mapping MRS representations to Discourse Representation Structures (DRS) (Klassen)
- Efficient HPSG generation for a non-configurational language (Crysmann, Packard)
- ERG Semantic Documentation (Bender, Oepen, Flickinger)

Grammar engineering for language documentation

- AGGREGATION: one year into pilot
  - Goals: automatically extract answers to Grammar Matrix questionnaire from IGT (ODIN, field projects)
  - Extracting underlying word order and case systems from 567 IGT (Bender et al at LaTeCH 2013)
    - Definition of Xigt - yet another xml format for IGT
    - Bender et al 2012 (COLING): Testing import from Toolbox lexicon for [ctn]
- Ling 567:
  - Field languages: Penobscot [aaq] (Quinn), Halkomelem [hur] (Gerdts),
  - Other: Frisian [frr], Bosnian-Serbo-Croatian [bcs], Lakota [lkt], Classical Chinese [lzh], Ryukuan [mvi], Yiddish [ydd]

New this year: MMT with ACE

- Faster system run times (though final run took 1h7m)
- More coverage (fewer system timeouts)
- Compatible with Condor (yay!)
  - => Possibility of parallelization (though not explored)
- Possibility of respecting ICONS representation of information structure

A big thank you to
Woodley & Sanghoun!

Items with end-to-end output

	aaq	eng	frr	hur	ita	lkt	lzh	mvi	ydd
aaq	6	6	6	6	6	5	7	6	5
eng	9	12	14	10	10	13	14	11	11
frr	8	11	16	8	10	15	12	10	12
hur	1	4	4	9	5	5	7	5	1
ita	8	10	11	7	11	10	9	8	9
lkt	7	9	12	9	9	12	11	8	9
lzh	8	10	11	12	9	11	14	10	10
mvi	9	11	10	9	9	9	11	12	8
ydd	8	11	14	8	9	13	12	10	13

Total number of outputs

	aaq	eng	frr	hur	ita	lkt	lzh	mvi	ydd
aaq	136466	68	224	408	240	65	76	66	4901
eng	92962	15	5305	3264	23	55	828	1238	522
frr	86943	82	231862	204	129	3210	148	1139	157
hur	4992	4	6	180	5	21	25	92	6
ita	83836	31	5255	3200	25	25	13	33	35190
lkt	61081	34	109	92	22	28	15	30	342
lzh	69013	334	579	588	360	427	494	1168	1180
mvi	71625	123	401	672	119	260	208	475	14476
ydd	98309	21	10527	78	15	65	138	1136	60

MatrixDoc: http://moin.delph-in.net/MatrixDocTop

- Primary authors: Antske Fokkens and Varya Gracheva
- Available for most existing libraries; should now be added for any additional libraries
- Contents (per library):
  - Description of options
  - Description of generated analyses
  - Tips & tricks
  - Citation information

Pozen (2013): *Lexical and Compositional Semantics for HPSG Parse Re-ranking*

- Use first (=most frequent in SemCor) WN sense for each lemma (given pos); Map to WN semantic file

- Mallet MaxEnt model; cast the problem as a discriminative classifier doing binary classification between preferred and dispreferred MRSs.

- Use classifier to rerank top-N output from syntactic MaxEnt parse selection model

- Features are synthetic features representing MRS predicate-argument triples (among others)

- WN annotated ERG parses over SemCor available: https://sites.google.com/site/zpozen/clms-thesis

Pozen (2013): *Lexical and Compositional Semantics for HPSG Parse Re-ranking*

	dev		test	
	Accuracy	ER	Accuracy	ER
No realpreds				
Top 1	42.17	3.36	45.38	2.86
Top 3	62.65	5.10	63.86	−1.12
Top 10	76.71	3.33	78.31	0.00
Random MRS baseline	13.06		12.63	
MRS model accuracy	30.52		34.54	
Lemma				
Top 1	43.37	5.37	42.17	−2.86
Top 3	65.06	11.22	61.85	−6.74
Top 10	77.91	8.33	78.31	0.00
Random MRS baseline	10.18		9.62	
MRS model accuracy	30.92		32.93	
SF				
Top 1	40.16	0.00	43.37	−0.71
Top 3	62.25	4.08	65.46	3.37
Top 10	77.51	6.67	79.92	7.41
Random MRS baseline	8.68		8.31	
MRS model accuracy	34.14		28.92	

Gracheva (2013): *Markers of Contrast in Russian: A Corpus-Based Study*

- Extract examples of ŽE and -TO (clitics marking contrast) from the Russian National Corpus

- Existing tests for contrastiveness often inapplicable

- -TO often marks a contrastive topic, with a contrastive focus elsewhere in the sentence (among other uses)

- -ŽE marks contrastive focus, including possibly of the whole sentence

- Challenges for analyzing info-structure in naturally occurring speech; subtlety of info-structure meanings

Schneider (in progress): The Effect on Deep Dependency Parsing from Training the Stanford

- Map DeepBank (Flickinger et al 2012) trees to PTB (Marcus et al 1993) format, including: node labels, tree geometry

- Train Stanford parser (Klein and Manning 2003) on exported DeepBank trees, and test with DDEC (Bender et al 2011)

- Hypothesis: More consistently annotated trees (from DeepBank) will lead better recovery of deep dependencies

- In progress, but preliminary results suggest a (narrowly) negative result: The Stanford parser's dependency labeling code is too closely tied to PTB trees.

(Towards) Declarative, Rule-Based Semantic Transfer (Slayden)

Declarative transfer

- Desiderata:
  - "One rule (set) to ring (include) them all."
    - both transfer directions should be supported by the same rule set
  - Maintainability/expressiveness
    - powerful organizing structures: to prevent rules from becoming too unwieldy
  - Performance
    - this is the most serious issue facing many DAG rewriting systems
    - constraining the problem by capitalizing on properties of linguistic inputs will be crucial

(Towards) Declarative, Rule-Based Semantic Transfer (Slayden)

Problems so far

- Opposing concerns:
  - For conceptual ease and authoring maintenance, rules want to be *small and distributed*
  - But in a in a declarative rewrite system it's not obvious how to allow *fragmentary rules to interact* with each other. i.e.
    - how do you "make reference" (grab on) to structures output by other rules—in *all* licensed ways
    - in general, the rewrite *system* (and not the rule author) has to solve this hard and expensive problem on the fly

And then there was this: