

Banks Meeting: PropBanks

2009-02-24

Martha Palmer

<http://www.bbn.com/NLP/OntoNotes>



PropBank - *press*

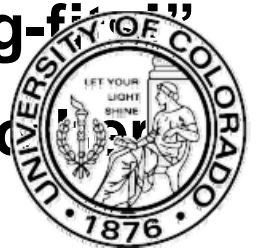


- *press* Roleset press.01 "apply pressure":
 - Verbnet: push 12.1, force 59.1
 - Roles:
 - Arg0: *presser*
 - Arg1: *pressed on*
 - Arg2: *goal, pressed for*
 - [ArgM-DIS *But*] [Arg0 *Ford, Jaguar's biggest shareholder since lifting its stake to 10.4% this week*], is *pressing* [ArgM-ADV *harder*] [Arg2 *for talks with Sir John.*]
- *press* Roleset press.02 "move":
 - Roles:
 - Arg0: *mover*
 - Arg1: *moved*
 - ArgM-DIR: *ahead, usually*
- WordNet – 14 verb senses

Functional tags for modifiers or adjuncts



- Variety of ArgM's:
 - TMP - when? *yesterday, 5pm on Saturday, recently*
 - LOC - where? *in the living room, on the newspaper*
 - DIR - where to/from? *down, from Antartica*
 - MNR - how? *quickly, with much enthusiasm*
 - PRP/CAU -why? *because ... , so that ...*
 - REC - himself, themselves, each other
 - ADV - hodge-podge, miscellaneous, “nothing-f...”
 - PRD - this argument refers to or modifies and



Adding Lemmas



- **New corpora often require Frame Files for previously unseen lemmas or edits to existing Frame Files**
- **Adding Frame Files – Currently 4,184 (5600+ senses)**

Data	New Frame Files	New Frame Sets	Edits
E Broadcast News	80	13	2
E Broadcast Conv	86	29	8
E/C Parallel TB	87	28	10
From VerbNet	79		
Total Added	344	70	20

English PropBank YR 3 Status



- **Corpora, Type coverage, 100%**
 - Lemmas: 4,184 (up from 3100 WSJ-Newswire)
 - Framesets: 5600+ (multiple subcategorization frames for a lemma)
- **Corpora, token coverage, 100%***
 - Y1 (NW, 300K) delivered
 - Y2 (BN, 200K) delivered – 33,800 instances completed
 - YR2 & YR3 (ECTB, 250K) – 30,000 instances completed
 - Y3 (BC, 200K) delivered – 27,600 instances, completed
 - Went quite a bit more slowly
 - Will also PropBank P2.5 (parallel MT evaluation data)
 - Y4 (NG/WL) Data being selected, with a goal of improving coverage for targeted lexical items
 - * *does not yet include YR 2/3 eventive nouns/nominalizations*

CPB: Milestones

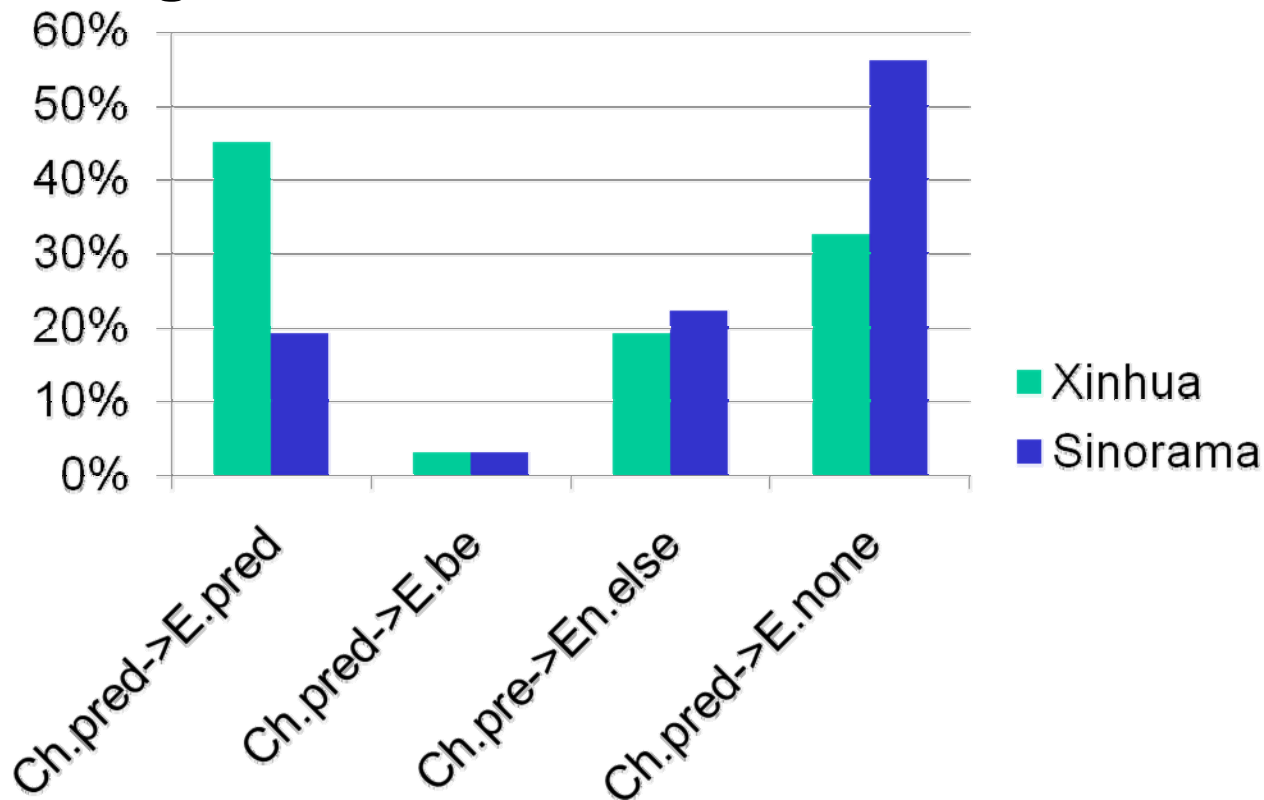


Version	CPB1.0	CPB2.0	CPB.xx	CPB.xx
Date	2005	2007	2008	2009
Words	250K	500K	800K	950K
Source	Xinhua	+Sinorama	+BN (300k)	+ BC (150K)
OntoNotes Data	partial	partial	yes	yes
Predicates	4,865	11,765	14,128	15,192
Framesets	5,298	12,555	15,077	??
Frameset/pred	1.09	1.06	1.07	??
Propositions	37,183	92,180	143,303	174,144

Using Parallel Propbanks to enhance word-alignments – Choi, Palmer, Xue



- **GIZA++ Alignments of Xinhua and Sinorama**



- **Use predicate-argument structures to verify and expand word-alignments for verb predicates**

- **Two approaches**
 - Top-down – start with GIZA++ matched predicates, *verify* by checking argument matches, use thresholds
 - Bottom-up – start with matching arguments, given enough matching arguments assume predicates match - *expansion*
- **Test with NIST MT eval data (thanks to ISI)**
 - 273 verb types, 100 sentences
- **Top-down approach effective with Sinorama**
 - GIZA++ generates 129 mappings, but w/ acc 57.76
 - T-D filters out 67 for 69 total, but acc goes up to 78.09
- **Bottom-up approach improves Xinhua**
 - Add 22 new mappings w/ acc 50.38
 - Thresholds filter out 4, acc goes up to 63.89