Brandeis University

Multilingual Propbanking

Nianwen Xue

7/15/2011

LSA Summer Institute

Deep semantics out of reach…

- Complete representation of real world knowledge = Natural Language Understanding?

NLU

  - Only build useful representations for small vocabularies

- Major impediment to accurate Machine Translation, Information Retrieval and Question Answering

“Shallow” semantic annotation

- Captures critical dependencies,
- Reflects clear, replicable, sense distinctions
- Based on a broad-coverage, informative, lexical resource
- Supports training of supervised automatic taggers
- Works for other languages

Syntactic variations

- The Congress increased the debt ceiling.
- The debt ceiling was increased by the Congress.
- The debt ceiling increased.
- The increase of the debt ceiling.

Accounting for syntactic variations: semantic roles

- The Congress increased the debt ceiling.
  ARG1
- The debt ceiling was increased by the Congress.
  ARG1
- The debt ceiling increased.
  ARG1
- The increase of the debt ceiling.
  ARG1

Semantic roles

- The Congress increased the debt ceiling.
  ARG1
- The debt ceiling was increased by the Congress.
  ARG1
- The debt ceiling increased.
  ARG1
- The increase of the debt ceiling.
  ARG1

ARG0: *causer of increase*
ARG1: *thing increasing*
ARG2: *amount increased*
ARG3: *start point*
ARG4: *end point*

Syntactic variations

- Greg broke the projector.
- The projector was broken by Greg.
- The projector broke into pieces.

Accounting for syntactic variations: semantic roles

ARG1

- Greg broke the projector.

ARG1

- The projector was broken by Greg.

ARG1

- The projector broke into pieces.

Semantic roles

ARG1

- Greg broke the projector.

ARG1

- The projector was broken by Greg.

ARG1

- The projector broke into pieces.

REL: BREAK:
ARG0: “breaker”
ARG1: “Thing broken”

Semantic role labels predicate-specific, but with trends…

- Arg0 = agent
- Arg1 = direct object / theme / patient
- Arg2 = indirect object / benefactive / instrument / attribute / end state
- Arg3 = start point / benefactive / instrument / attribute
- Arg4 = end point

Consistency for **Arg0** and **Arg1**, not so much for Arg2, Arg3

Define roles for…

- Expect, inhibit
- Replace, translate, give
- Rise
- Send

expect

Roles:

Arg0: expecter

Arg1: thing expected

Example: Transitive, active:

Portfolio managers ***expect*** *further declines in interest rates.*

Arg0:	*Portfolio managers*
REL:	***expect***
Arg1:	*further declines in interest rates*

inhibit

Roles:

Arg0: inhibitor

Arg1: process inhibited

Example: Transitive, active:

Prostacyclin inhibits platelet aggregation, smooth muscle cell proliferation, and vasoconstriction.

Arg0:	*Prostacyclin*
REL:	*inhibit*
Arg1:	*platelet aggregation, smooth muscle cell proliferation, and vasoconstriction*

replace

Roles:

Arg0: replacer

Arg1: old thing

Arg2: new thing

Example: *Continental Air replaced its top executive for the sixth time in as many years.*

Arg0:	*Continental Air*
REL:	*replaced*
Arg1:	*its top executive*
ArgM-TMP:	*for the sixth time in as many years.*

translate

Roles:

Arg0: agent/translater

Arg1: new thing

Arg2: old thing

Example:

Thus, Acj6 and Drifter translate PN lineage information into distinct dendritic targeting specificity.

Arg0:	*Acj6 and Drifter*
REL:	*translate*
Arg1:	*distinct dendritic targeting specificity*
Arg2:	*PN lineage information*

give

Roles:

Arg0: giver

Arg1: thing given

Arg2: entity given to

Example: double object

The executives gave *the chefs a standing ovation.*

Arg0:	*The executives*
REL:	*gave*
Arg2:	*the chefs*
Arg1:	*a standing ovation*

rise

Roles (no ARG0 for unaccusative verbs)

- Arg1 = Logical subject, patient, thing rising
- Arg2 = EXT, amount risen
- Arg3* = start point
- Arg4 = end point

Sales rose 4% to $3.28 billion from $3.16 billion.

The Nasdaq composite index added 1.01 to 456.6 on paltry volume.

Verbs are given six arguments

- Bump, burst, dilute, inch, knock, nudge, pop, send, shoot, sink

ARG0: causer of motion
ARG1: entity in motion
ARG2: extent of motion
ARG3: starting point
ARG4: end point
ARG5: direction

Define the roles for “leave”…

Word Senses in PropBank

- Ignoring word sense not feasible for 700+ verbs
  - *Mary left the room*
  - *Mary left her daughter-in-law her pearls in her will*

Frameset leave.01 "move away from":
Arg0: entity leaving
Arg1: place left

Frameset leave.02 "give":
Arg0: giver
Arg1: thing given
Arg2: beneficiary

ArgM's: not all arguments are predicate-specifc

- TMP - when?
- LOC - where at?
- DIR - where to?
- MNR - how?
- PRP - why?
- REC - himself, themselves, each other
- PRD - this argument refers to or modifies another
- ADV – others

Stuff swept under the rug…

- Negation
- Auxiliary

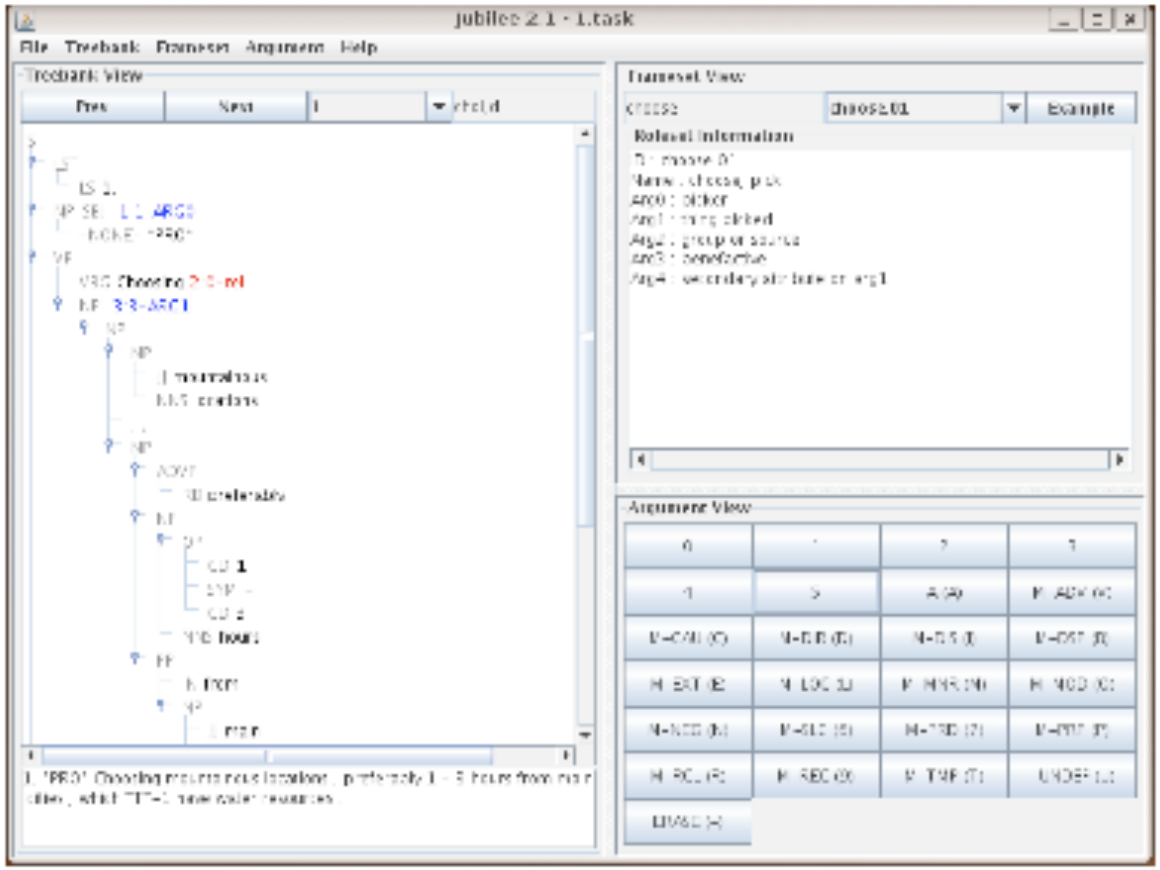
The cow [**ARGM-AUX** did] [**ARGM-NEG** not] eat grass with a fork.

Blanco and Moldovan 2011
"Semantic representation of negation using focus"

PropBank annotation procedure

- PTB II – Extract all sentences of a verb
- Create Frame File for that verb
- First pass: Automatic tagging
- Second pass: Double blind hand correction
  - Inter-annotator agreement 84%, vital to machine learning
- Third pass: *Solomonization* (adjudication)
- Validation and release

Jubilee



http://code.google.com/p/propbank

Measure ITA

- At argument level
- At predicate level

Defining Framesets (1)

- Defining framesets involves characterizing the arguments of a verb in terms of (a) their syntactic realizations (subcat frames) and (b) their "semantic" properties.
- Two subcat frames are the same if they have the same type and number of arguments, otherwise they are different
- One subcat frame subsumes another if the arguments of the latter is a subset of the former.
- All subcat frames that belong to a frameset should either be identical to or subsume one another.

Defining Frameset (2)

- Syntactic realizations and semantic properties are expected to coincide most of the time: difference (similarity) in meaning is reflected in difference (similarity) in syntactic realizations (c.f. Levin 1993)

Theoretic basis

- Beth Levin, *English Verb Classes and Alternations* (1993)
- *"Behavior of a verb . . . is to a large extent determined by its meaning"* (p. 1)

  Amanda hacked the wood with an ax.

  Amanda hacked at the wood with an ax.

  Craig notched the wood with an ax.

  *Craig notched at the wood with an ax.
- Can we move from syntactic behavior back to semantics?

Levin classes *(Levin, 1993)*

- 3100 verbs, 47 top level classes, 193 second and third level
- Each class has a syntactic signature based on alternations.

  John broke the jar. / The jar broke. / Jars break easily.

  *John cut the bread. / *The bread cut. / Bread cuts easily.*

  *John hit the wall. / *The wall hit. / *Walls hit easily.*

Propbank as a linguistic resource

Frame files: a computational lexicon of how the arguments **CAN** be realized.
Corpus: Many examples of how the arguments are **ACTUALLY** realized.

[ARG0 The special persecutor] [PRED is conducting] [ARG1 an investigation into the governor's handling of the senate appointment]

[ARG0 The special persecutor] [SUPPORT is conducting] [PRED an investigation] [ARG1 into the governor's handling of the senate appointment]

Available at http://verbs.colorado.edu

Propbank as an engineering benchmark

- CoNLL-04, 05, 08, 09 Shared Task
- Xue and Palmer, *EMNLP04*
- Pradhan, et. al., *NAACL04, ICDM03,*
- Sardeneau, et. al, *ACL03,*
- Chen & Rambow, *EMNLP03,* Gildea & Hockemaier, *EMNLP03*
- Gildea & Jurafsky, *CL02,* Gildea & Palmer, *ACL02*
- Yi, Loper and Palmer, *NACCL07*

Linguistics/Engineering tradeoffs

- Arguments can only be annotated within the same clause
  - Can arguments ever be found outside of a clause?
- Semantic role added to syntactic constituents
  - One constituent per argument?
- Numbered argument or linguistically motivated semantic roles
  - Consistency and replicability
- Granularity of Propbank senses

Sentence-external arguments?

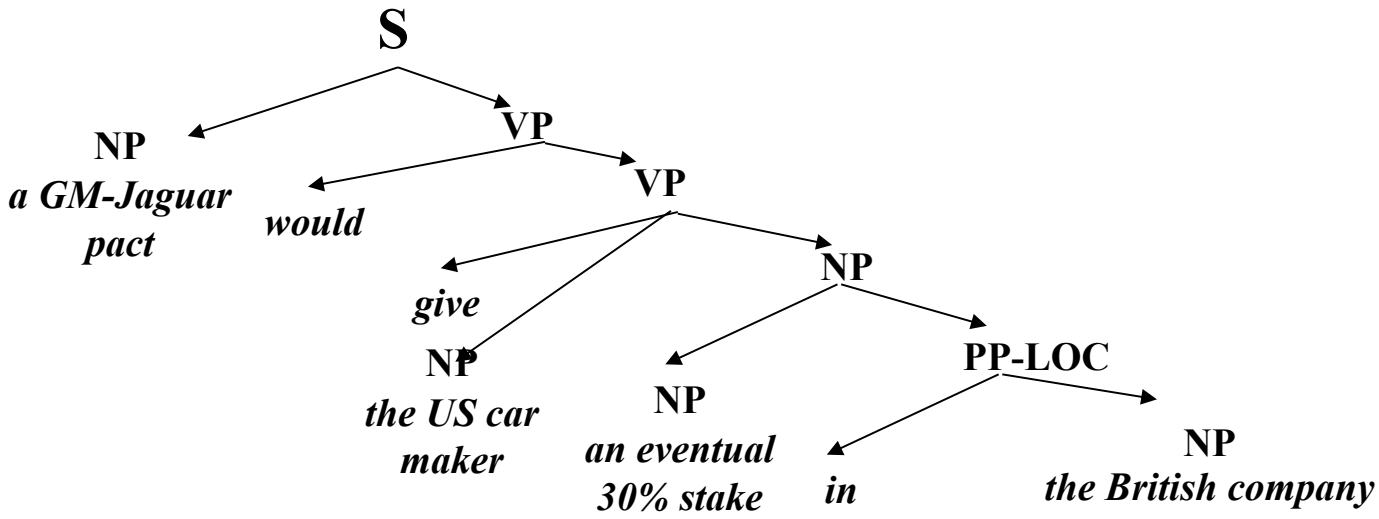
[**ARG0** The two companies] [**PRED** produce] [**ARG1** market pulp, containerboard and white paper]. The goods could be manufactured closer to customers, saving [PRED shipping] Costs.

Gerber and Chai 2010

A TreeBanked sentence

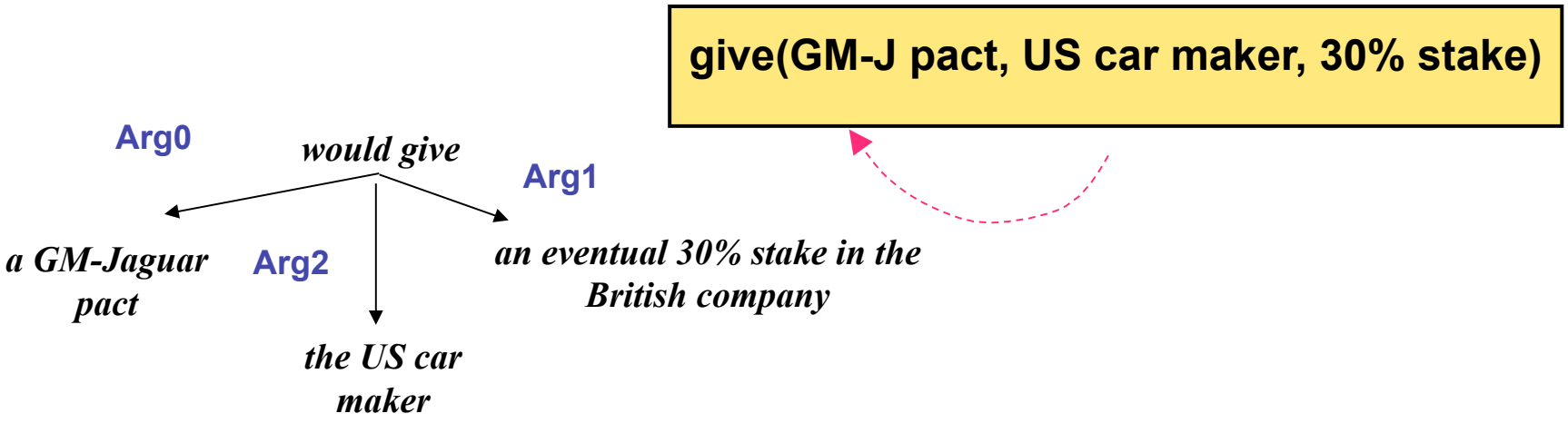
Marcus, et al, 1993

A GM-Jaguar pact would give the U.S. car maker an eventual 30% stake in the British company.



The same phrase, PropBanked

A GM-Jaguar pact would give the U.S. car maker an eventual 30% stake in the British company.



Multiword Expressions

- Multiword Expressions (MWEs) are common in natural language and easily mastered by native speakers

Idiomatic Expressions

"Stop pulling my leg!"

Verb Particle Constructions

"You must get over that shyness."

Light Verb Constructions

"I took a walk down by the sea"

Other Complex Predicates

नकर् गया 'nikal gayaa' (lit. "exit-went")
meaning: left or departed

Multiword Expressions

- MWEs' are challenging due to their semantic flexibility:
  - "You must **get over** that shyness." [Partially compositional]
  - "My baby **threw up** all over me!" [Non-compositional]

- The subcateogorization frame of the predicate is no longer soley dependent on the verb alone:
  - "My baby **threw up** all over me!"
    ➔ THROW UP(AGENT)
  - *"My baby **threw** all over me!"
    ➔ THROW(AGENT,PATIENT)

Multiword Expressions

- Natural language processing applications have to a large degree ignored them and usually fail to distinguish between purely compositional interpretations and interpretations that take multiword expressions into account.

- Lexical resources associated with semantic role labeling tasks such as PropBank (Palmer et al., 2005), VerbNet (Kipper et al., 2008) and FrameNet (Johnson, et. al., 2001) include a fair amount of verb particle constructions, but do not have a general, portable treatment of light verb constructions.

PropBank Annotation of LVCs

- Focus on the PropBank annotation of the Light Verb Constructions (LVCs)
- Goals:
  - Define and label LVCs in a consistent and effective manner
  - Develop a schema that will be suitable for all PropBank languages: English, Arabic, Chinese, and Hindi
  - Explore practical challenges involved in the cross-linguistic analysis of LVCs

Light Verb Constructions

- Linguistically accepted as a type of complex predicates
- Many studies from differing frameworks have characterized complex predicates as a fusion of two or more predicative elements (Rosen 1997, Goldberg 1993, Mohanan 1997, Alsina 1997, Butt 1997)
- Our approach was adapted from Butt 2004.

Light Verb Constructions (Butt 2004)

- Light verb + predicating complement (true predicate)
- Light verb is semantically bleached; does not hold full predicating power
- Light verb + true predicate combine to predicate as a single semantic element
- LVC can be paraphrased with the verbal form of the true predicate without loss of core meaning

I took a walk down by the sea.

'I walked down by the sea.'

Light Verb Constructions

- True Predicates:
  - Nouns: *Found in all four languages*
    - "take a walk", "give a lecture"
  - Adjectives: *Found in Hindi*
    - मुझे तुम अच्छे लगे
      'to-me you nice seem'
      "I like you"
  - Verb: *Found in Hindi*
    - मैंने सब कुछ कर लया
      'I everything do took'
      "I've done everything"
  - Object NP of the preposition: *Found in Arabic*
    - سأقوم بزيارة سيدنا إلياس
      conduct.I PREP-visit our.St Ilias"
      ""I will visit St. Illias

Final Annotation Scheme

Pass 1: Light Verb Identification

- Predicating expression is labeled ARG-PRX

Pass 2: LVC Annotation

- Arguments and modifiers of the LVCs are annotated

Pass 3: Deterministic relation merge

- Arguments and modifiers are taken from Pass 2

Example:

"John took a brisk walk through the park."

```
REL: took
ARG-PRX:
     a brisk walk
```

```
ARG0: John
REL: walk
ARG-MNR: brisk
ARG-DIR:
  through the park
```


```
ARG0: John
REL: took walk
ARG-MNR: brisk
ARG-DIR:
  through the park
```

Distinguishing LVCs from MWEs

- Definition of LVC (Revisit): based on Butt (2004)
  - semantically bleached light verb
  - predicating element within the direct object of the light verb
  - semantics of light verb + true predicate is distinct from either element taken independently
  - syntax often distinct from expected subcategorization frame of the verb
- Allows annotators to make reasonably consistent judgments of LVCs

Distinguishing LVCs from MWEs

- LVCs are semi-productive constructions existing in a continuum from fixed idiomatic expression to purely compositional collocation
- Borderline cases exist in all languages:
  - make an exception
  - take charge
- LVC or idiomatic expression?
  - Most English speakers no longer use verbal counterparts ‘except’ or ‘charge (with a responsibility)’ but LVCs are entrenched

Distinguishing LVCs from heavy usages

- Several verbs seem to participate in complex predication but contribute in varying levels to semantics:
  - light: produce an alteration ‘alter’
  - light: issue a complaint ‘complain’
  - heavy: register a complaint
- No deterministic method for measuring extent to which verb contributes to semantics, or to distinguish borderline verbs from normal, heavy usages

Types of semantic role labels:

Kevin broke the LCD projector.

break (agent(Kevin), patient(LCD-projector))

Fillmore, 68

cause(agent(Kevin),
broken(LCD-projector))

Jackendoff, 72

agent(A) -> intentional(A), sentient(A),
causer(A), affector(A)
patient(P) -> affected(P), change(P),…

Dowty, 91

Could have gone “deeper”

ARG1→THEME

- Greg broke the projector.

ARG1→THEME

- The projector was broken by Greg.

ARG1→THEME

- The projector broke into pieces.

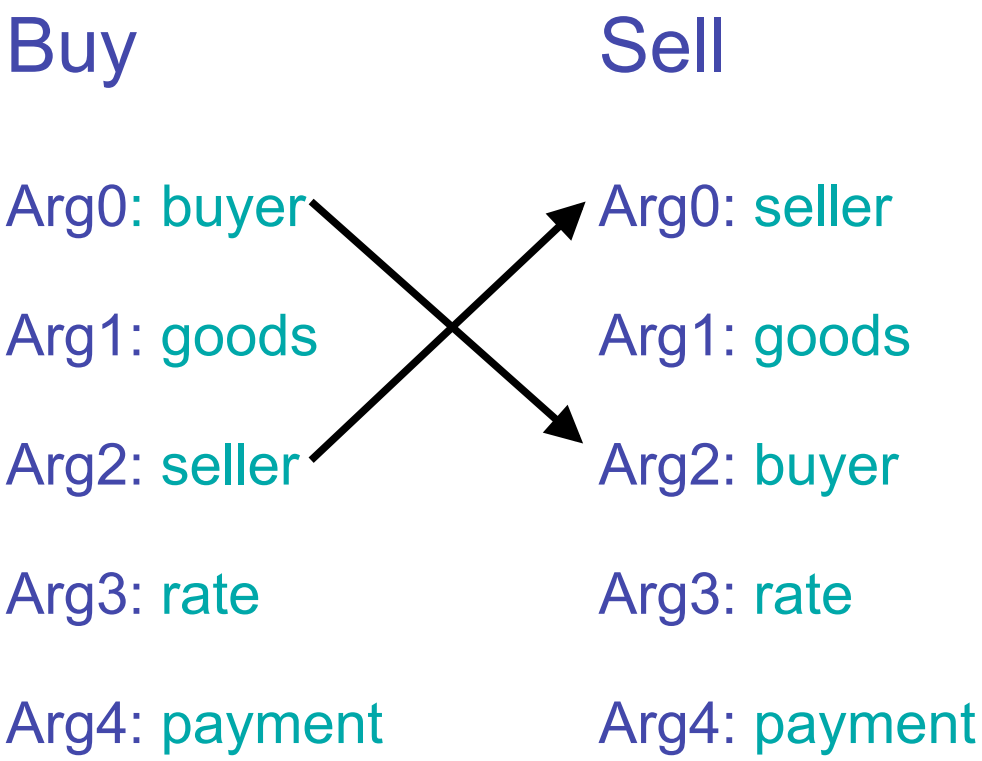
AGENT, PATIENT,SOURCE, GOAL, THEME, SOURCE, BENEFICIARY, …

But consistency might have taken a hit…

Mapping from PropBank to VerbNet

Frameset id = leave.02	Sense = give	VerbNet class = future-having 13.3
Arg0	Giver	Agent
Arg1	Thing given	Theme
Arg2	Benefactive	Recipient

PropBank/FrameNet



PropBank semantic roles for core arguments are more predicate-specific

Granularity of word senses: WordNet – Princeton

(*Miller 1985, Fellbaum 1998)*

On-line lexical reference (dictionary)

- Nouns, verbs, adjectives, and adverbs grouped into synonym sets
- Other relations include hypernyms (ISA), antonyms, meronyms
- Typical top nodes - 3 out of 25
  - *(act, action, activity)*
  - *(animal, fauna)*
  - *(artifact)*

WordNet – call, 28 senses

1. name, call -- (assign a specified, proper name to;
   "They named their son David"; …)
   -> LABEL
2. call, telephone, call up, phone, ring -- (get or try to get into communication (with someone) by telephone;
   "I tried to call you all night"; …)
   ->TELECOMMUNICATE
3. call -- (ascribe a quality to or give a name of a common noun that reflects a quality; *"He called me a bastard"; …)*
   -> LABEL
4. call, send for -- (order, request, or command to come;
   "She was called into the director's office"; "Call the police!")
   -> ORDER

WordNet: - call, 28 senses

WN2 , WN13,WN12 WN15 WN26

WN3 WN19 WN4 WN 7 WN8 WN9

WN1 WN22

WN20 WN25

WN18 WN27

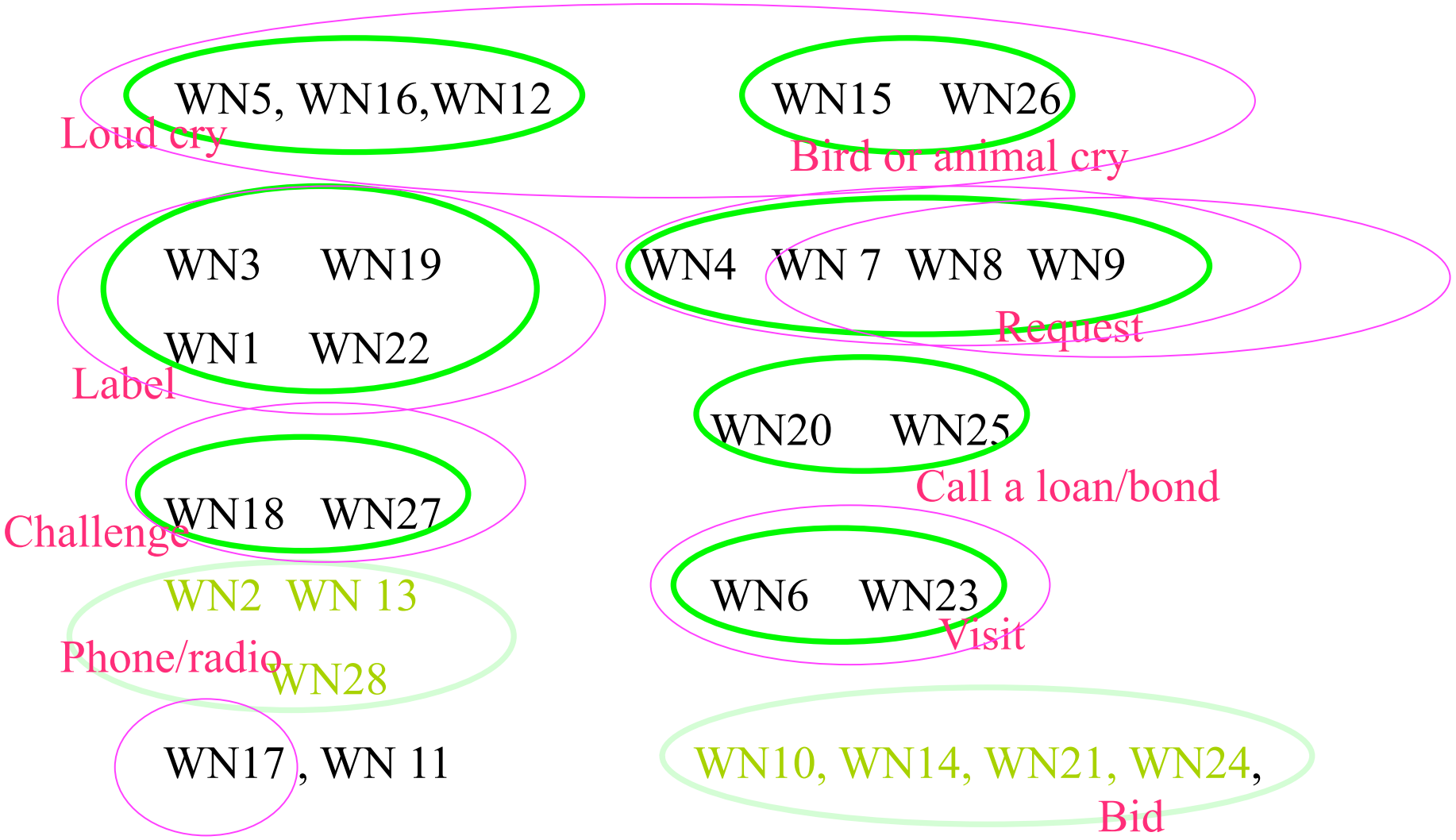
WN5 WN 16 WN6 WN23

WN28

WN17 , WN 11 WN10, WN14, WN21,WN24

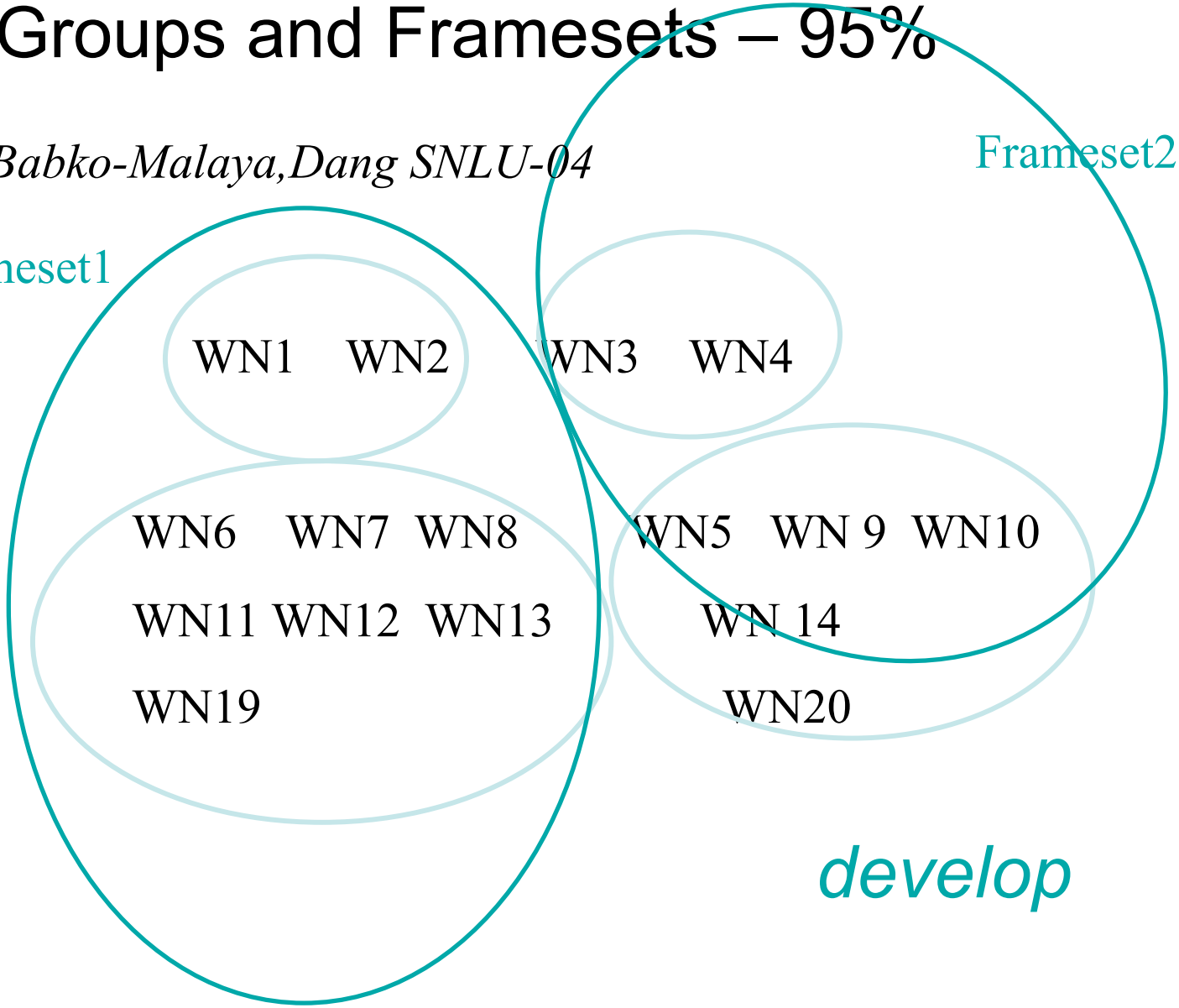
WordNet: - call, 28 senses, Senseval2 groups (engineering!)

Overlap with PropBank Framesets



Overlap between Senseval2 Groups and Framesets – 95%

Palmer, Babko-Malaya,Dang SNLU-04



Multilingual Propbanking

- Chinese Propbank
  - Xue 2008, CL, Xue and Palmer 2008, NLE
- Korean Propbank
  - Palmer et al, 2006, LDC
- Arabic Propbank
  - Palmer et al, 2008, LREC
- AnCora, for Spanish and Catlan
  - Taulé et al, 2008, LREC

A Chinese Treebank Sentence

国会/Congress 最近/recently 通过/pass 了/ASP 银行法/banking law

"The Congress passed the banking law recently."

```
(IP (NP-SBJ  (NN 国会/Congress))
    (VP (ADVP (ADV 最近/recently))
        (VP (VV 通过/pass)
            (AS 了/ASP)
            (NP-OBJ (NN 银行法/banking law)))))
```

The same sentence, propbanked

国会/Congress 最近/recently 通过/pass 了/ASP
银行法/banking law
“The Congress passed the banking law recently.”

REL:通过/pass
Arg0 (agent):国会/Congress
Arg1 (theme):银行法/banking law
ArgM-TMP:最近/recently

Machine Translation

- 他/he 在/at 这/this 个/CL 文件/document 上/on 签/sign 了/ASP 自己/self 的/DE 名字/name
  SYSTRAN: He has signed own name in this document
  Correct: He signed his own name on this document

- 他/he 在/at 这/this 个/CL 文件/document 上/on 签字/sign
  SYSTRAN: He signs in this document
  Correct: He signed this document.

- Problem: Prepositional phrase is NOT semantic adjunct.

MT: Further examples

- 俄罗斯/Russia 撤回/withdraw 军队/army.
  SYSTRAN: Russia withdraws the army.
  Correct: Russia withdrew the army.

- 俄罗斯/Russia 军队/army 撤回/withdraw 莫斯科/Moscow.
  SYSTRAN: The Russian army withdraws Moscow
  Correct: The Russian army withdrew to Moscow.

- Problem: Argument is the goal (arg2), not theme (arg1)!

Arguments, Frames, Framesets

火车/train 正在/now 通过/pass 遂道/tunnel
"The train is passing through the tunnel."

火车/train 正在/now 通过/pass
"The train is passing."

Frameset1

国会/Congress 最近/recently 通过/pass 了/ASP
银行法/banking law
"The Congress passed the banking law recently."

银行法/banking law 最近/recently 通过/pass 了/ASP
"The banking law passed recently."

Frameset2

Chinese Propbank as a linguistic resource

[**ARG0** 特别 检察官] [**ARG1-a** 对 州长 的 参议员 任命 事件]
special persecutor into governor DE senate appointment
[**PRED** 进行] [**ARG1-b** 调查]
conduct investigation

[**ARG0** 特别 检察官] [**ARG1** 对 州长 的 参议员 任命 事件]
special persecutor into governor DE senate appointment
[**SUPPORT** 进行] [**PRED** 调查]
conduct investigation

[ARG0 The special persecutor] [PRED is conducting] [ARG1 an investigation into the governor's handling of the senate appointment]

[ARG0 The special persecutor] [SUPPORT is conducting] [PRED an investigation] [ARG1 into the governor's handling of the senate appointment]

Availabe at http://www.ldc.upenn.edu

Summary

- "Meaning" is shallow semantic annotation that captures critical dependencies, semantic role labels and sense distinctions
- Supports training of accurate, supervised automatic taggers
- Methodology ports readily to other languages
  - English PropBank release – spring 2004
  - Chinese PropBank release – fall 2004
  - Korean PropBank release – summer 2005

References

- Martha Palmer, Dan Gildea, Paul Kingsbury. 2005.The Proposition Bank: A Corpus Annotated with Semantic Roles, Computational Linguistics Journal, 31:1

- Nianwen Xue, Martha Palmer. 2009. Adding semantic roles to the Chinese Treebank. Natural Language Engineering 15(1): 143-172

- Jena Hwang, Archna Bhatia, Clare Bonial, Aous Mansouri, Ashwini Vaidya, Nianwen Xue and Martha Palmer. Propbank annotation of multilingual light verb constructions. In Proceedings of the Fourth Linguistic Annotation Workshop. Uppsala, Sweden.