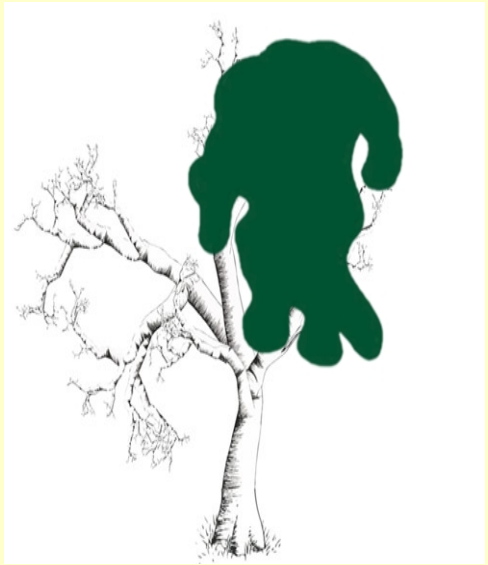
GLARF and the 2nd Stage of Parsing

Combining Parsing, SRL, NE tagging, Temporal Tagging, ...



CUNY-NLP Seminar

Adam Meyers, NYU

March 25, 2011

Collaborators

- Michiko Kosaka (Monmouth)
- Nianwen Xue (Brandeis)
- Shasha Liao (NYU)
- Heng Ji (CUNY)
- Ralph Grishman (NYU)
- Yusuke Shinyama (formerly of NYU)
- Others at NYU Proteus Group

Outline

- **Approaches to Combining Annotation**
- Introduction to GLARF
- Merging into one GLARF-based Theory
- Some GLARF results
- Future Work

Terminology

- Linguistic Annotation: A formal description of linguist properties of a text
- Automatic Annotation
  - Examples: parser output, NE tagging, other taggers
- Manual Annotation
  - Examples: Penn treebank, Manual NE labels, TimeBank and other “Banks”, etc.
- Transducer
  - System for automatically creating annotation

Approach 1: Annotating Multiple Linguistic Phenomena within a Single Theoretical Framework

- All annotation must share some assumptions: POS, segmentation, tokenization, headedness (if possible), constituents (if possible), etc.
- Possibly augment/modify theory over time
- Possibly revise previous annotation schemes
- Disadvantages:
  - Theory (unless revised) may overly constrain annotation of some phenomena
  - Limits input to those willing to work in that framework
- Examples:
  - Tübingen Treebank of Written German
  - Czech Dependency Treebank
  - Kyoto Corpus Treebank
  - Copenhagen Dependency Treebank

Approach 2: Merging Annotation A La Carte

- Given a set of annotations, convert each annotation into a common ***physical form,*** typically ***c***haracter offset-based XML
- Possible to incorporate the work of many different research projects (with no theory in common)
- Disadvantage: Glosses over incompatibilities between annotations (segmentation, tokenization, constituents, headedness, etc.) which may make cross-phenomenon generalizations difficult.
- Examples:
  - *Ontonotes: The 90% Solution* (Hovy, et. al. 2006)
  - *MASC: the Manually Annotated Sub-Corpus of American English* (Ide, et. al. 2008)
  - *Combining Independent Syntactic and Semantic Annotation Schemes* (Verhagen, et. al. 2007)

Our Approach: Merging A la Carte Annotation while Changing it in Our Own Biased Way

- **Advantages:**
  - Resulting annotation fits together (like Approach 1), sharing segmentation, tokenization, headedness (if possible), constituents (if possible)
  - Incorporates annotation undertaken by several research projects which assume different theories (like Approach 2)
  - Corrects some annotation errors through conflict resolution
- **Disadvantages**
  - Our theory may be inappropriate for some phenomena (like Approach 1)
  - Some information may be lost/mishandled during translation
- **Similar to CONLL 2008/2009 Shared Task**
  - GLARF is more ambitious (a possibly more difficult shared task)
  - GLARF's Logic1 “level” that is easier to generate automatically and covers all words in the sentence.

Outline

- Approaches to Combining Annotation
- **Introduction to GLARF**
  - **Grammatical and**
    Logical
    Argument
    Representation
    Framework
- Merging into one GLARF-based Theory
- Some GLARF results
- Future Work

Why GLARF?

- Annotation of different linguistic phenomena are incompatible with each other.
- It is difficult to create NLP applications that use evidence from representations of several phenomena.
- It is difficult to correlate disparate types of linguistic evidence from corpora.
- GLARF provides one way of solving these problems

GLARF is:

- Grammatical and Logical Argument Representation Framework
  - A framework for representing linguistic information
    - A GLARF-based theory
  - Any theory within that framework
    - In GLARF, we assume...
  - A GLARF representation of a sentence or phrase
    - In the GLARF of sentence X, ...
- A system for automatically producing GLARF
  - Available for Download
  - Most Common Feedback from Users:
    - It is better than I expected it would be
  - This suggest that a Good tag line for GLARF might be:
    - GLARF, It's Better Than You Think it is

Introduction to GLARF

- Languages: English, Chinese, Japanese
- Typed Feature Structure: Maximal Information
- Multiple Dependency Tuples: Less Info + headedness assumptions
- Produces a single-theory analysis
  - Not 100% Reversible
- GLARF System combines:
  - hand-annotation
  - automatically generated annotation
  - combination of manual/automatic annotation
- GLARF approach to merging annotation was part of the NSF-funded ULA (Unified Linguistic Annotation) project involving: The Penn Treebank, PropBank, NomBank, The Penn Discourse Treebank, TimeBank and the Pittsburg Opinion TreeBank

2 Main Purposes for GLARF

- The Second Stage of a 2-stage (LFG-style) parser – the first stage is a standard tree-bank-based parse (PTB, Chinese PTB, Kyoto Corpus)
  - Other automatic output (at least NE) is incorporated
  - Before PTB-based parsers, 2 stage parsers (Hobbs and Grishman 1976) were popular.
    - 1st stage = Syntax
    - 2nd stage = Regularized (fill gaps, transform passives, etc.)
- A merging program for manual/automatic annotation (plus additional info derived via rules, dictionary information, etc.).
- These 2 functions are indistinguishable from each other.

Example Sentence

- ***Afterwards, she decided to perform the operation.***
  - Current Sentence (Sentence Number 1) Offset of first character = 29
  - Previous Sentence (Sentence Number 0) is ***The doctor ran some tests***
- ***PDTB (and TimeML):***
  - ***Afterwards***: ARG1 = previous S, ARG2 = current S
- ***PropBank***
  - ***decided***: ARG0 = ***she***, ARG1= ***to perform the operation***, ARGM-TMP = ***Afterwards***
  - ***perform***: ARG0 = ***she***, ARG1= ***the operation***
- ***NomBank***
  - ***operation: ARG0 = she, Support = perform***
- **Penn Treebank**
  - **(S (ADVP (RB Afterwards)) (, ,)**
    (NP (PRP she))
    (VP (VBN decided)
    (S (VP (TO to)
    (VP (VB perform)
    (NP (DT the) (NN operation))))))) (. .))

GLARF TFS

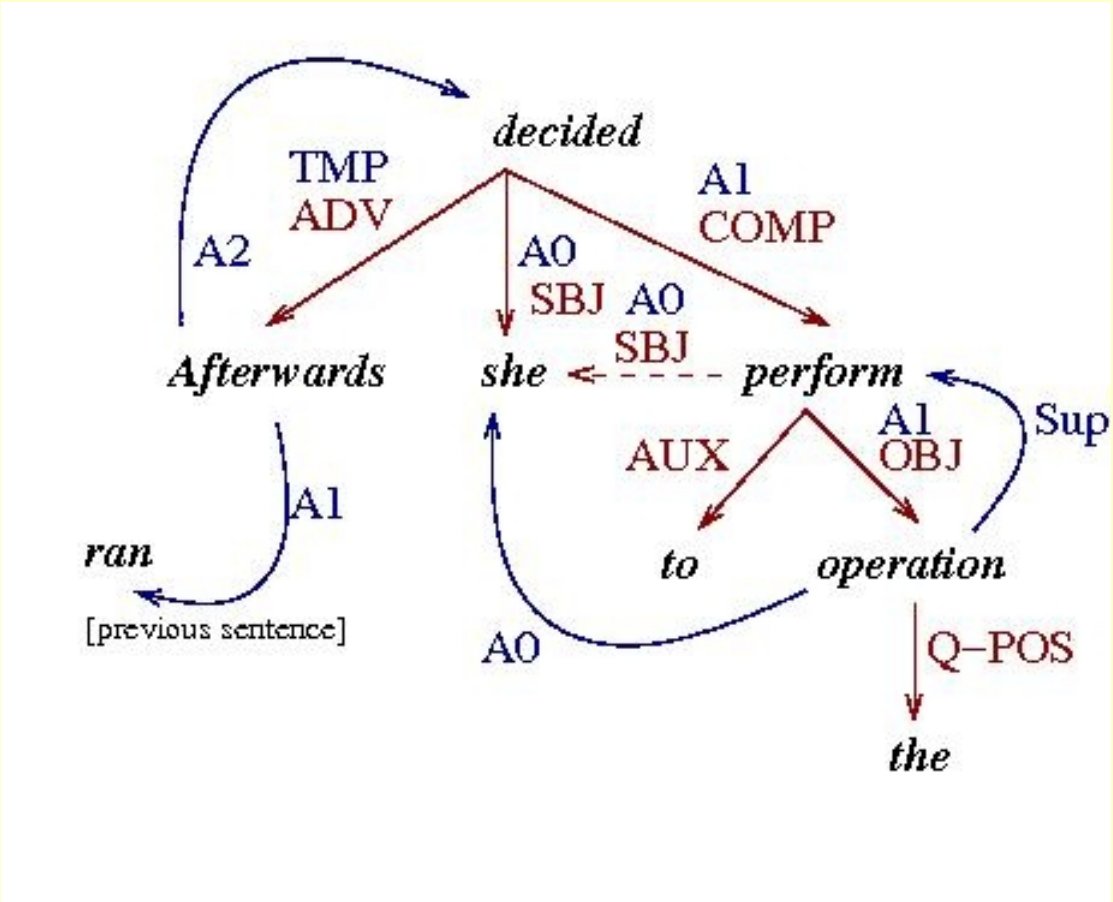
(S (**ADV** (ADVP (**HEAD** (ADVX (**HEAD** (RB ***Afterwards*** 0))
(**P-ARG1** (S (EC-TYPE PB) (INDEX 0+0))
(**P-ARG2** (S (EC-TYPE PB) (INDEX 0))
(RELATION-TYPE AFTER)))
(INDEX 1) (POINTER 0:1))))
(PUNCTUATION (, **,** 1))
(**SBJ** (NP (**HEAD** (PRP ***she*** 2)) (INDEX 2) (POINTER 2:1))))
(**PRD** (VP (**HEAD** (VG (**HEAD** (VBD ***decided*** 3))
(**P-ARG0** (NP (EC-TYPE PB) (INDEX 2)))
(**P-ARG1** (S (EC-TYPE PB) (INDEX 5)))
(**P-ARGM-TMP** (ADVP (EC-TYPE PB) (INDEX 1)))
(SEM-TENSE PAST)))
(**COMP** (S (**L-SBJ** (NP (EC-TYPE INF) (INDEX 2)))
(**PRD** (VP (**HEAD** (VG (**AUX** (TO ***to*** 4))
(**HEAD** (VB ***perform*** 5))
(**P-ARG0** (NP (EC-TYPE PB) (INDEX 2)))
(**P-ARG1** (NP (EC-TYPE PB) (INDEX 4)))
(INDEX 3)))
(**OBJ** (NP (**Q-POS** (DT ***the*** 6))
(**HEAD** (NX (**HEAD** (NN ***operation*** 7)))
(**P-SUPPORT** (VG (EC-TYPE PB) (INDEX 3)))
(**P-ARG0** (NP (EC-TYPE PB) (INDEX 2)))))
(INDEX 4) (POINTER 6:1)))
(PB-POINTER 4:1)))
(POINTER 4:2) (INDEX 5)))
(POINTER 3:1)))
(PUNCTUATION (. **.** 8)) (POINTER 0:2) (TREE-NUM 1) (INDEX 0))

GLARF Dependency Tuples (Abbreviated)

Logic1	Surface	Logic2	Functor	Off	POS +	Argument	Off	POS +
ADV	ADV	TMP	*decided*	45	VBD PAST	*Afterwards*	29	RB TMP
SBJ	SBJ	ARG0	*decided*	45	VBD PAST	*she*	41	PRP
COMP	COMP	ARG1	*decided*	45	VBD PAST	*perform*	56	VB
SBJ	NIL	ARG0	*perform*	56	VB	*she*	41	PRP
OBJ	OBJ	ARG1	*perform*	56	VB	*operation*	68	NN
AUX	AUX	NIL	*perform*	56	VB	*to*	53	TO
Q-POS	Q-POS	NIL	*operation*	68	NN	*the*	64	DT
NIL	NIL	ARG0	*operation*	68	NN	*she*	41	PRP
NIL	NIL	SUPP	*operation*	68	NN	*perform*	56	VB
NIL	NIL	ARG1	*Afterwards*	29	RB TMP	*ran*	11	VBD PAST
NIL	NIL	ARG2	*Afterwards*	29	RB TMP	*decided*	45	VBD PAST

Triple Dependency Graph

- Solid Red Lines
  - surface/L1
  - surface/L1/L2
- Dashed Red Lines
  - L1/L2
  - L1 only
- Solid Blue Lines
  - L2 only
- Red Labels
  - Surface/L1
- Blue Labels
  - L2



Logic1 and Surface Dependencies

- Surface dependencies form a tree
  - like S-Structure or C-Structure
- Logic1 dependencies form a directed acyclic graph
  - like F-Structure or D-Structure (if Empty Category = Antecedent)
- Many Logic1 and Surface Dependencies are the same
- Logic1/Surface Distinction represents syntactic regularization and gap filling
  - Passive, Relative Clause Gaps, Subjects of Infinitives, VP Deletion, etc.
- Apparent cycles are removed via Logic1/Surface distinction
  - PARENTHETICAL and RELATIVE are Surface dependencies, but their gaps represent logic1 dependencies
    - *Mary, [John believed __], was a vampire*
    - *I want [the book [that John was reading __]]*

Logic 1 Role Labels regularize relations between predicate/argument pairs

- **Red:** Predicate
- **Blue:** Logic1 OBJ
- **Yellow:** Logic1 SBJ

They were eaten by the giant clam

≈

The giant clam ate them

Logic 2 Dependencies

- Logic2 dependencies form a directed graph with cycles
- Logic2 includes argument relations that do not fit neatly into the Surface vs Logic1 dichotomy.
- Includes semantics-based argument relations that are in complementary distribution, because the functors belong to distinct parts of speech:
  - Arguments of Verbs: PropBank
  - Arguments of (subset of) nouns: NomBank
  - Arguments of (subset of) adverbs, prepositions, coord/subord conjunctions: overt PDTB. TimeML TLINK and analagous relations with NP arguments

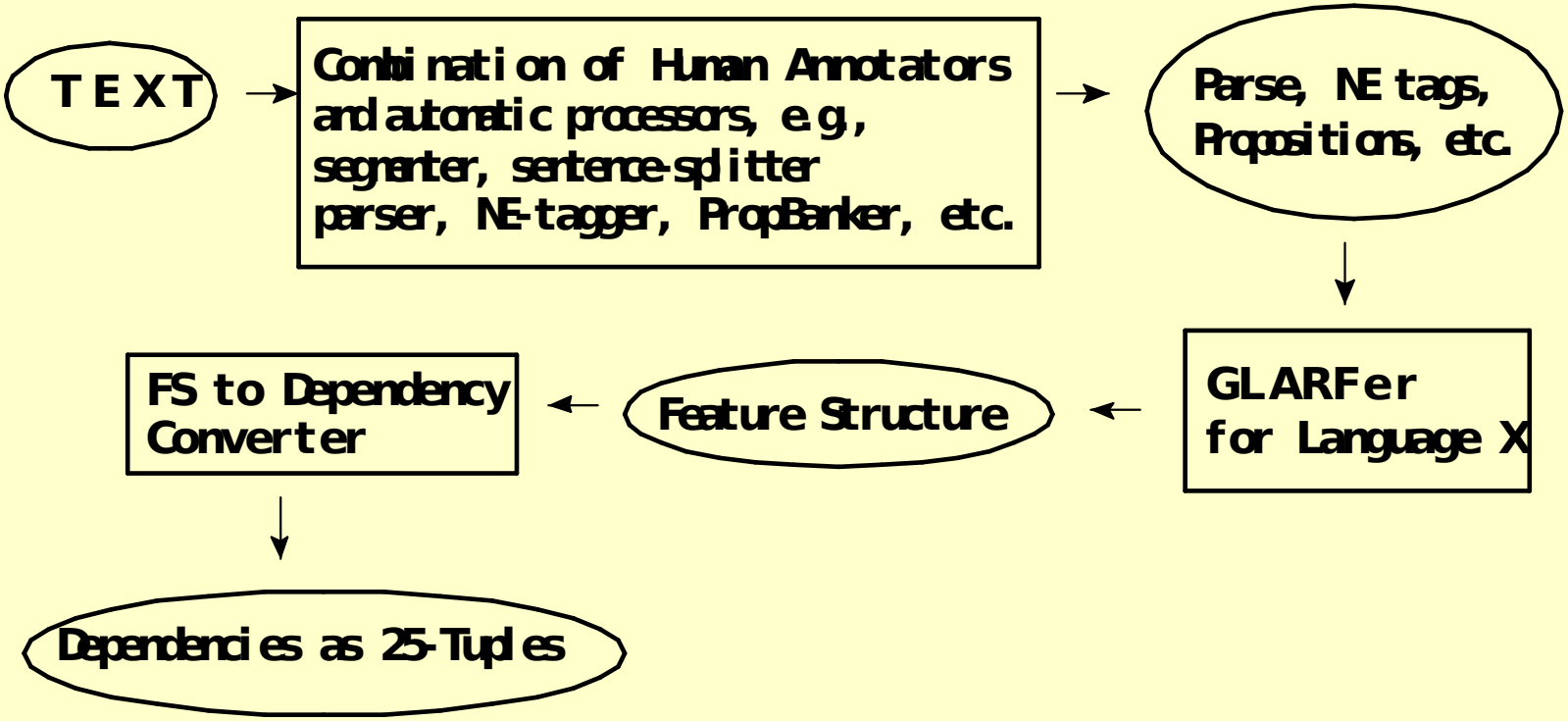
Full GLARF TFS and Tuples

- More detail: morphology, semantic classes, senses from PropBank/NomBank, etc.
- Regularizes across productive syntactic regularities, distinguishing logical and surface SBJ/OBJ, e.g., passive, relative clause, etc.
- Regularizes Conjunctions, distinguishing the functor (conjunction) from the conjuncts, the latter acting like heads for purposes of tuples
- Incorporates recognition of Named Entities, Time Expressions, Numbers, and similar phenomena.
- Handles non-headed constructions (multi-word expressions, range phrases, the-more-the-merrier constructions, etc.)
- Handles degree/comparative/superlative complements
- Current Research in MT, time sequencing/causation relations among events, times and other elements
- Current tuples are 25-tuples including base forms, senses, etc.

Outline

- Approaches to Combining Annotation
- Introduction to GLARF
- **Merging into one GLARF-based Theory**
- Some GLARF results
- Future Work

Annotation Merging with GLARF



Current (English) System

- Expected Input:
  - Sentence-split input with offsets
  - Named entity input for ACE classes (GPE, PER, FAC, LOC, ORG)
    - Automatic annotation (JET) or BBN's hand annotation of Penn Treebank
    - Syntactic Tree: PTB-parse (Charniak) or manual treebank (PTB)
- Manually created rules add additional information to all input
  - Syntactic Regularizations (Logic1)
  - Error Correction (Part of Speech, Constituent Structure)
  - Special constructions: Time/Number Expressions, Legal Cases, etc.
  - TimeML information
- Can incorporate input annotation or produce as part of GLARF system:
  - PropBank, NomBank, Overt PDTB relations
- Cascade of filters starting with parse tree and assuming strict rule ordering

Merging Considerations

- When are 2 relations part of the same “level”?
- When can 2 annotations be assumed to represent different parts of a single relation?
- What if 2 annotations assume Different Constituent Structures?
- What if merging causes undesirable dependency structures, e.g., loops?

Justification for Levels

- Surface and Logic1 levels
  - Sentence-internal regularizations based on consensus of popular theories
  - Passive, standard gap filling constructions (WH, relative, parenthetical, control, raising, VP-deletion), and a other phenomena compatible with surface/logic1 distinction
  - Distinction prevents loops in Logic1
    - Modifier relations are SURFACE if they contain gaps that modify the containing structure (parentheticals, relatives)
- The Level Logic2
  - Phenomena that don't neatly fit into Logic1
  - Phenomena are compatible with each other
    - Mergeable or in Complementary Distribution
  - Do not prevent loops or constrain to sentence-internal phenomena, etc.

Logic2: PropBank/NomBank

- PropBank includes verb alternations that may not be syntactic
  - *The pilot*/**ARG0** *flew the plane*/**ARG1**
  - *The plane*/**ARG1** *flew*.
- NomBank clusters arguments due to related verbs and other factors
  - *Rome's*/**ARG1** *destruction by tourists*/**ARG0**
  - *John's*/**ARG0** *capacity for understanding*/**ARG1**
- PropBank/NomBank argument relations are finer grained than Logic1

Logic2: Overt PDTB, TimeML/TLINK, Extensions

- PDTB relations and TimeML TLINK relations are not compatible with sentence-internal Surface/Logic1 relations
- When a TLINK signal and a PDTB predicate are the same, they also have the same arguments and their information is mergeable
  - *The doctor ran some tests.*/**ARG1** *Afterwards*, *she decided to perform the operation.*/**ARG2**
  - **ARG1 = EventInstance, ARG2 = relatedToEventInstance**
- Other times, they are in complementary distribution
  - *The test was performed*/**ARG2**, *but I don't know the results*/**ARG1**. (ONLY PDTB)
  - *She left*/***ARG1*** *on Tuesday*/**ARG2** (Only TimeML)
- Natural Extentions of analysis, e.g., to non-time-related PPs
  - *She left*/**ARG1** *because of the problem*/**ARG2**

Segment/Token/Constituent Compatibility Assumptions

1. Non parse-tree units are compatible with parse-tree units if there are no “crossing boundaries”. There are 2 subcases:
   a) they correspond to parse-tree units OR
   b) they can be analyzed as evidence sub-constituents
2. When 1 is not possible, the difference is predictable by rules or heuristics

Adding Subconstituents

- Nombank
  - (NP cotton and acetate fibers) →
    (NP (NP cotton and acetate) fibers)
  - (NP a Thursday night practice) →
    (NP a (NP Thursday night) practice)
- BBN NEs
  - (NP New York-based Loews Corp.) →
    (NP (ADJP (NP New York) - based) (NP Loews Corp.))
  - (NP Republican Rudolph Guliani 's) →
    (NP (NP Republican) (NP Rudolph Guliani) 's)

Resolving Token Level Conflicts

- BBN Named Entities
  - *(ADJP (NNP New) (JJ York-based))* →
    *(ADJP (**STEM** (NP New York))*
    (PUNCTUATION (HYPH -))
    *(**HEAD** (VBN based)))*
  - *(NP (JJ U.S.-Japanese) (NNS relations))* →
    (NP (N-POS (NP (CONJ1 (NP U.S.))
    (CONJUNCTION (HYPH -))
    (CONJ2 (NP Japanese))))
    (HEAD (NNS relations)))
- NomBank
  - higher/**ARG1** student/**ARG0**-test/**ARG2** scores
  - (NP **(A-POS** (ADJP (HEAD (JJR higher 9)) (INDEX 1)))
    (**N-POS** (NP (**N-POS** (NX (HEAD (NN student 10.1)) (INDEX 2))
    (PUNCTUATION (HYPH - 10.2))
    (**HEAD** (NX (HEAD (NN test 10.3)) (INDEX 3)))))
    (HEAD (NX (**HEAD** (NN scores))
    (P-ARG1 (NP (EC-TYPE PB) (INDEX 1)))
    (**P-ARG0** (NP (EC-TYPE PB) (INDEX 2)))
    (**P-ARG2** (NP (EC-TYPE PB) (INDEX 3))))))
- PTB ↔ Text alignment problems (I found 15 cases)
  - Predictable misalignments between text and hand-coded trees – ***cannot*** → ***can*** + ***not***, ***tis*** → **-t + is**
  - Rare Errors deletions, unpredictable textual changes

Logic2: Formal Difficulties

- Cycles resulting from interactions between predicate types
  - NomBank Support Verbs and PropBank Arguments
    - NomBank: *Mary*/ARG0 *took*/Support *a walk*
    - PropBank: *Mary*/ARG0 *took a walk*/ARG1
  - PDTB predicates are PropBank modifiers
    - PDTB: *Afterwards*, *she slept*/ARG2.
    - PropBank: *Afterwards*/ARGM-TMP, *she*/ARG0 *slept*
  - These cycles seem unavoidable without a Surface2/Logic2 distinction
- Predicates embedded inside their arguments
  - *The cow*/ARG1, *John*/ARG0 *said*, *jumped over the moon*/ARG1.
  - The ARG1 is a discontinuous constituent (perhaps irrelevant for dependency representation)
  - This is unavoidable – Further discussion on next few slides

Discontinous Arguments are interupted by a Self-Phrase Containing the Predicate

- Self-Phrase **P'**
  - Given: Predicate **P,** Argument **A,** Adverbial **P'**
    - **P'** is a PP, ADVP, parenthetical, etc.
  - Where **P'** is a child of **A**, **P'** is an ancestor of **P**
  - Convention: Listing **A** as the argument will be understood to mean **A minus P'**
- PropBank Parentheticals
  - (S-1 (NP The cow)
    (PRN (S (NP John)
    (VP (VBD said)
    (SBAR (-NONE- 0) (S –NONE- *T*-1)))))
    (VP jumped over the moon))
  - ARG1 of ***said*** = ***The cow*** + ***jumped over the moon***
- NomBank:
  - (S (NP The legislation) (PP at their request) (VP was (VP introduced (ADVP (RB early)))))
  - ARG1 of ***request*** = ***The legislation*** + ***was*** + ***introduced early***
- PDTB
  - (S (NP the company) (ADVP also) (VP disclosed what it did))
  - ARG2 of ***also*** = ***the company*** + ***disclosed what it did***

Adjunction Rule Unites Contiguous Clausal arguments

- (S (SBAR although
  (S preliminary findings suggest X))
  (NP the latest results)
  (VP suggest Y)) →
  (S (SBAR although
  (S preliminary findings suggest X))
  (S (NP the latest results)
  (VP suggest Y)))
- Rule allows ***although*** to select whole Ss as arguments
- Preserves dependency structure
- Defensible Constituent Structure

Eliminating apparent SBJ + VP discontinuities using PTB empty categories

- If an S contains an EC bound to an external NP, an argument need not include that NP
  - $John_i$ seems (S e_i to leave)
    - If ARG1 of ***seem*** is ***John*** + ***to leave***
    - ***John*** can be deleted for purposes of LOGIC2
- If PDTB argument of a coordinate conjunction is a VP plus its SBJ, add an EC to the VP and delete the SBJ
  - ***(S They try to (VP (VP watch the other ropes) and thus (VP time their pulls)))***
  - Args of ***thus: They*** + ***time their pulls*** and ***They*** + ***watch the other ropes***

Outline

- Approaches to Combining Annotation
- Introduction to GLARF
- Merging into one GLARF-based Theory
- **Some GLARF results**
- Future Work

GLARF Evaluation Description

- **Evaluate exact match of 4 out of 25-tuples**
  - **Logic1 Role Label, +/-Transparent, Functor, Argument**
  - 4 English: 46-100 sentences or 450 to 1500 tuples
    - 2 Written (WSJ and LET) and 2 Spoken (TEL and NAR)
  - CTB & KYO: 20 sentences or 400 to 600 tuples
- **+/- Transparent refers to whether or not a functor is semantically empty**
  - Conjunctions (*and/o*r), Transparent Nouns (*variety of birds*), copulas, etc.
- **F-score (F-T) also calculated ignoring transparency**
  - Reduces Number Correct
    - precision = correct/output-length
    - recall = correct/answerkey-length
- **Answer Keys: different for tb/parse system output**
  - Genuine Ambiguity
  - Features measured not always specified as correct/incorrect
  - Sometimes more than one way to represent same concept in framework

Evaluation on Test Corpora

- English: WSJ = Wall Street Journal, LET = correspondence, TEL = telephone transcripts, NAR = transcripts of narratives (LET, TEL, NAR are from OANC)
- Chinese: CTB = Chinese Treebank
- Japanese: KYO = Kyoto Corpus
- The Chinese/Japanese systems less developed compared to English
- English Evaluations: 46-100 sentences, Chinese/Japanese: 20 sentences (number of relations = 400-1500)

Treebank					Parser			
ID	Prec	Rec	F	F-T	Prec	Rec	F	F-T
WSJ	83.0%	84.2%	83.6%	87.1%	80.2%	78.9%	79.5%	81.8%
LET	92.9%	92.3%	92.6%	93.3%	89.9%	85.9%	87.8%	87.8%
TEL	76.2%	81.2%	78.6%	82.2%	74.8%	74.5%	74.7%	77.4%
NAR	89.7%	84.0%	82.3%	84.1%	75.7%	74.7%	75.2%	76.1%
CTB	87.8%	89.1%	88.4%	88.7%	87.3%	80.4%	83.7%	83.7%
KYO	91.3%	91.0%	91.1%	91.1%	84.9%	86.2%	85.5%	87.8%

More Evaluations

- Number of sentences: 50 En News, 46 En Blog, 53 Ch News, 40 Ja News
- Parser Output Only
-

Corpus	Precision	Recall	F-Score
En NEWS (JENAAD)	731/815=89.7%	715/812=90.0%	89.9%
En BLOG	704/844=83.4%	704/899=78.3%	80.8%
Ch Nwire	1031/1415=72.9%	1031/1352=76.3%	74.5%
Ja NEWS (JENAAD)	764/843=91.0%	764/840=90.6%	90.8%

2 Indirect Evaluations of Automatically Generated GLARF

- Improved Giza++ score for 2010 MT research at NYU
  - Automatically aligned English/Chinese Logic1 GLARF graphs
  - Derived mappings to reordered English text to be like Chinese
  - Lowered Alignment Error Rate from 51.9% (raw text) to 50.6% on Test Corpus (1505 sentences hand-aligned by LDC released GALE Y1 Q4)
-

NYU ACE Event 12/2005 on DEV-TEST	**VDR**	**VMD**
Chunker + SBJ/OBJ Heuristics	18.3	20.9
Parser + SBJ/OBJ/Passive Heur.	21.3	24.8
GLARF Logic1	25.8	31.2
GLARF Logic1 + Logic2	27.6	32.8

Other Work Using GLARF

- Part of 5-W System for 2009 SRA team GALE
  - NYU's part of an ensemble system for answering: *Who*, *What*, *Why*, *Where* & *How* questions
  - Parton, et. al. (2009), Yaman, et. al. (2009)
- Several IE systems used GLARF-based patterns
  - NYU dissertations: Zhao (2005), Shinyama (2007)
- Creating Data: CONLL 2008 & 2009 English Shared Task
  - Automatic NP-internal relations from GLARF (Prec: 83.9%-88%)
  - Automatic Split Tokenization due to hyphens (Prec: 85.5%-92.2%)
  - NomBank dependencies (filtered through GLARF)
  - Surdeanu, et. al. (2008) and Hajič, et. al. (2009)

Outline

- Approaches to Combining Annotation
- Introduction to GLARF
- Merging into one GLARF-based Theory
- Some GLARF results
- **Current and Future Work**

Alpha Version of GLARF for Download

- English only
- Packaged with version of Charniak parser and JET NE tagger/sentence splitter
- Intended for automatically created annotation
- Open Source, except for encrypted version of Comlex Syntax (due to LDC license)
- Free for Non-Profit and Research Use
- Commercial inquiries are welcome
- http://nlp.cs.nyu.edu/meyers/GLARF.html

What's Next?

- Work on Causation and Temporal Relations
  - Automatic system influenced by TimeML and PDTB specifications
  - Incorporates more PPs and NPs
- Further work on Machine Translation
  - Expanding using Chinese/Modified-English with MOSES translation system
- Online System
  - Functionality, Feedback and Collaboration

Summary

- GLARF is A framework for cutting out a theory from a merger of several different annotation schemata
- Our transducer derives the analysis from:
  - Manual and/or automatic annotation
  - An ordered series of filters
- A Theoretically-biased merger provides consistent structures for use in applications
  - We believe its consistency outweighs negative impact of our biases
- GLARF has been used successfully as part of many systems at NYU
- English GLARF has been 10 years in the making and is now available for download

Extra Slides

- Some Explanations
  - Multiple Correct Answers
  - Transparency Clarification
- Chinese and Japanese Examples
- Slides about contribution to CONLL

Multiple Correct Answers

Ambiguity	Corp	Treebank	Parser
1. Tokenization	NAR	*2 - hour, 2 - cent*	2-hour, 2-cent
2. Prefix?	KYO	大 */big* + 枠 */framework*	大枠 */the big picture*
3. Encoding of zero	CTB	二０００年 */year 2000*	二 000 年 */year 2000*
4. Attachment (relative)	LET	*thousands [of people] [who face obstacles]*	*thousands of [people [who face obstacles]]*
5. Conj Scope	TEL	*[pearls or [beads of some sort of necklace]]*	*[[pearls or beads] of some sort of necklace]*
6. Mod ambiguity 多種多様な */varied* + 事業 */businesses*	KYO	Relative Clause *businesses that are varied*	Adjectival Modifier *various businesses*
7. POS ambiguity 进口 /export = N or V	CTB	进口五十亿 *Exportation of 5 billion*	进口 五十亿 *Exported 5 billion*

Transparency Explanation

- Transparency: conjunctions, partitives, light Vs, copulas
  - Arguments act like semantic head(s)
  - ***[John and Mary] ate [a bag of sandwiches]***
    - red = functor of NP, yellow = semantic heads

Chinese: 汉语中，关联词和被动句也有很明显的特点。

In Chinese, conjunctions and passive sentences also have very obvious features.

Surf	L1	L2	Func	Arg
ADV	ADV		有/have	中/in
SBJ	SBJ	A0	有/have	和/and
ADV	ADV		有/have	也/also
OBJ	OBJ	A1	有/have	特点/features
OBJ	OBJ		中/in	汉语/Chinese
CONJ	*CONJ		和/and	关联词/conjunctions
CONJ	*CONJ		和/and	被动句/passive sentences
A-POS	A-POS		特点/features	的/DE
COMP	*COMP		的/DE	明显/obvious
ADV	ADV		明显/obvious	很/very

生命・財産を守ることは国家の責務だ。

It is the state's duty to protect lives and assets.

L1	Surf	L2	Func	Arg
***PRD**	PRD		だ /is	責務 /duty
	SBJ		だ /is	こと /fact
SBJ			責務 /duty	こと /fact
COMP	COMP		責務 /duty	国家 /state
PRT	PRT		国家	の
COMP	COMP		こと /fact	守る /protect
PRT	PRT		こと	は
OBJ	OBJ		守る /protect	NULL-CONJ
***CONJ**	CONJ		NULL-CONJ	財産 /assets
PRT	PRT		財産	を
***CONJ**	CONJ		NULL-CONJ	生命 /lives

CONLL Splitting at Hyphens/Slashes 1

- Split tokens:
  - Assign POS tags
    - Automatic results for sample of 179 tokens
      - 153 correct (85.5%), 14 incorrect (7.8%), 12 unclear (6.7%)
  - Decimal token numbers
- (VP (NP (NNP New 6)
  - (NNP York 7.1)))
  - (HYPH – 7.2)
  - (VBN based 7.3))

NP-internal Relations

- NP internal relations used for CONLL
  - Title: ***Mr.*** *John Smith*
  - Post-Hon: *John Smith* ***Jr.*** *III*, *Inc.*, *Ph.D.*, *etc.*
  - APPOsite: *John Smith*, ***president of the U.S.***
  - SUFFIX: *John* ***'s***
  - Near 100% accuracy for small sample
    - 45 correct, 2 unclear
- All NP GLARF Roles
  - RELATIVE, COMP, A-POS, T-POS, Q-POS, etc.
  - 224 correct (83.9%), 32 wrong (12%), 11 unclear (4.1%)

CONLL Splitting at Hyphens/Slashes 2

- Split Segments iff:
  - COMLEX words, numbers, prefixes (from a list)
  - Required byBBN NE tags (we made a gazatteer)
- Relations from GLARF
  - Conjunction cases: ***Japan-U.S.*** *agreement*
  - Everything else (distinguish HMOD/HEAD)
    - GLARF distinguishes them further