

Transducing Logical Relations from Automatic and Manual GLARF

Adam Meyers[†], Michiko Kosaka[‡], Heng Ji^{*}, Nianwen Xue[◇],

Mary Harper[▽], Ang Sun[†], Wei Xu[†] and Shasha Liao[†]

[†]New York Univ., [‡]Monmouth Univ., [◇]Brandeis Univ., ^{*}City Univ. of New York, [▽]Johns Hopkins Human Lang. Tech. Ctr. of Excellence & U. of Maryland, College Park

Abstract

GLARF relations are generated from treebank and parses for English, Chinese and Japanese. Our evaluation of system output for these input types requires consideration of multiple correct answers.¹

1 Introduction

Systems, such as treebank-based parsers (Charniak, 2001; Collins, 1999) and semantic role labelers (Gildea and Jurafsky, 2002; Xue, 2008), are trained and tested on hand-annotated data. Evaluation is based on differences between system output and test data. Other systems use these programs to perform tasks unrelated to the original annotation. For example, participating systems in CONLL (Surdeanu et al., 2008; Hajič et al., 2009), ACE and GALE tasks merged the results of several processors (parsers, named entity recognizers, etc.) not initially designed for the task at hand. This paper discusses differences between hand-annotated data and automatically generated data with respect to our GLARFers, systems for generating Grammatical and Logical Representation Framework (GLARF) for English, Chinese and Japanese sentences. The paper describes GLARF (Meyers et al., 2001; Meyers et al., 2009) and GLARFers and compares GLARF produced from treebank and parses.

2 GLARF

Figure 1 includes simplified GLARF analyses for English, Chinese and Japanese sentences. For each sentence, a GLARFer constructs both a Feature Structure (FS) representing a constituency analysis and a set of 31-tuples, each representing

up to three dependency relations between pairs of words. Due to space limitations, we will focus on the 6 fields of the 31-tuple represented in Figure 1. These include: (1) a functor (**func**); (2) the depending argument (**Arg**); (3) a surface (**Surf**) label based on the position in the parse tree with no regularizations; (4) a logic1 label (**L1**) for a relation that reflects grammar-based regularizations of the surface level. This marks relations for filling gaps in relative clauses or missing infinitival subjects, represents passives as paraphrases as actives, etc. While the general framework supports many regularizations, the relations actually represented depends on the implemented grammar, e.g., our current grammar of English regularizes across passives and relative clauses, but our grammars of Japanese and Chinese do not currently.; (5) a logic2 label (**L2**) for Chinese and English, which represents PropBank, NomBank and Penn Discourse Treebank relations; and (6) Asterisks (*) indicate *transparent* relations, relations where the functor inherits semantic properties of certain special arguments (*CONJ, *OBJ, *PRD, *COMP).

Figure 1 contains several transparent relations. The interpretation of the *CONJ relations in the Japanese example, include not only that the nouns [*zaisan*] (*assets*) and [*seimei*] (*lives*) are conjoined, but also that these two nouns, together form the object of the Japanese verb [*mamoru*] (*protect*). Thus, for example, semantic selection patterns should treat these nouns as possible objects for this verb. Transparent relations may serve to neutralize some of the problematic cases of attachment ambiguity. For example, in the English sentence *A number of phrases with modifiers are not ambiguous*, there is a transparent *COMP relation between *numbers* and *of* and a transparent *OBJ relation between *of* and *phrases*. Thus, high attachment of the PP *with modifiers*, would have the same interpretation as low attachment since *phrases* is the underlying head of *number of*

¹Support includes: NSF IIS-0534700 & IIS-0534325 Structure Alignment-based MT; DARPA HR0011-06-C-0023 & HR0011-06-C-0023; CUNY REP & GRTI Program. This work does not necessarily reflect views of sponsors.

1. English: A number of phrases with modifiers are not ambiguous				
Surf	L1	L2	Func	Arg
SBJ		A1	are	number
ADV	*ADV	NEG	are	not
PRD	*PRD	A2	are	ambiguous
	SBJ		ambiguous	number
Q-POS	Q-POS		number	a
COMP	*COMP	A1	number	of
WITH	*WITH		number	with
OBJ	*OBJ		of	phrases
OBJ	OBJ		with	modifiers
2. 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.				
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
3. Japanese: 生命・財産を守ることは国家の責務だ。 (The fact of) protecting lives and assets is the state's duty.				
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

Figure 1: GLARF 5-tuples for 3 languages

phrases. In this same example, the adverb *not* can be attached to either the copula *are* or the predicative adjective, with no discernible difference in meaning—this factor is indicated by the transparent designation of the relations where the copula is a functor. Transparent features also provide us with a simple way of handling certain function words, such as the Chinese word *De* which inherits the function of its underlying head, connecting a variety of such modifiers to head nouns (an adjective in the Chinese example.). For conjunction cases, the number of underlying relations would multiply, e.g., *Mary and John bought and sold stock* would (underlyingly) have four subject relations derived by pairing each of the underlying subject nouns *Mary* and *John* with each of the underlying main predicate verbs *bought* and *sold*.

3 Automatic vs. Manual Annotation

Apart from accuracy, there are several other ways that automatic and manual annotation differs. For

Penn-treebank (PTB) parsing, for example, most parsers (not all) leave out function tags and empty categories. Consistency is an important goal for manual annotation for many reasons including: (1) in the absence of a clear correct answer, consistency helps clarify measures of annotation quality (inter-annotator agreement scores); and (2) consistent annotation is better training data for machine learning. Thus, annotation specifications use defaults to ensure the consistent handling of spurious ambiguity. For example, given a sentence like *I bought three acres of land in California*, the PP *in California* can be attached to either *acres* or *land* with no difference in meaning. While annotation guidelines may direct a human annotator to prefer, for example, high attachment, systems output may have other preferences, e.g., the probability that *land* is modified by a PP (headed by *in*) versus the probability that *acres* can be so modified.

Even if the manual annotation for a particular corpus is consistent when it comes to other factors such as tokenization or part of speech, developers of parsers sometimes change these guidelines to suit their needs. For example, users of the Charniak parser (Charniak, 2001) should add the AUX category to the PTB parts of speech and adjust their systems to account for the conversion of the word *ain't* into the tokens *IS* and *n't*. Similarly, tokenization decisions with respect to hyphens vary among different versions of the Penn Treebank, as well as different parsers based on these treebanks. Thus if a system uses multiple parsers, such differences must be accounted for. Differences that are not important for a particular application should be ignored (e.g., by merging alternative analyses). For example, in the case of spurious attachment ambiguity, a system may need to either accept both as right answers or derive a common representation for both. Of course, many of the particular problems that result from spurious ambiguity can be accounted for in hind sight. Nevertheless, it is precisely this lack of a controlled environment which adds elements of spurious ambiguity. Using new processors or training on new treebanks can bring new instances of spurious ambiguity.

4 Experiments and Evaluation

We ran GLARFers on both manually created treebanks and automatically produced parses for English, Chinese and Japanese. For each corpus, we created one or more answer keys by correcting

system output. For this paper, we evaluate solely on the logic1 relations (the second column in figure 1.) Figure 2 lists our results for all three languages, based on treebank and parser input.

As in (Meyers et al., 2009), we generated 4-tuples consisting of the following for each dependency: (A) the logic1 label (SBJ, OBJ, etc.), (B) its transparency (True or False), (C) The functor (a single word or a named entity); and (D) the argument (a single word or a named entity). In the case of conjunction where there was no lexical conjunction word, we used either punctuation (commas or semi-colons) or the placeholder *NULL*. We then corrected these results by hand to produce the answer key—an answer was correct if all four members of the tuple were correct and incorrect otherwise. Table 2 provides the **Precision**, **Recall** and **F-scores** for our output. The **F-T** columns indicates a modified F-score derived by ignoring the +/-Transparent distinction (resulting changes in precision, recall and F-score are the same).

For English and Japanese, an expert native speaking linguist corrected the output. For Chinese, several native speaking computational linguists shared the task. By checking compatibility of the answer keys with outputs derived from different sources (parser, treebank), we could detect errors and inconsistencies. We processed the following corpora. English: 86 sentence article (wsj_2300) from the Wall Street Journal PTB test corpus (WSJ); 46 sentence letter from Good Will (LET), the first 100 sentences of a switchboard telephone transcript (TEL) and the first 100 sentences of a narrative from the Charlotte Narrative and Conversation (NAR). These samples are taken from the PTB WSJ Corpus and the SIGANN shared subcorpus of the OANC. The filenames are: 110CYL067, NapierDianne and sw2014. Chinese: a 20 sentence sample of text from the Penn Chinese Treebank (CTB) (Xue et al., 2005). Japanese: 20 sentences from the Kyoto Corpus (KYO) (Kurohashi and Nagao, 1998)

5 Running the GLARFer Programs

We use Charniak, UMD and KNP parsers (Charniak, 2001; Huang and Harper, 2009; Kurohashi and Nagao, 1998), JET Named Entity tagger (Grishman et al., 2005; Ji and Grishman, 2006) and other resources in conjunction with language-specific GLARFers that incorporate hand-written rules to convert output of these processors into

a final representation, including logic1 structure, the focus of this paper. English GLARFer rules use Comlex (Macleod et al., 1998a) and the various NomBank lexicons (<http://nlp.cs.nyu.edu/meyers/nombank/>) for lexical lookup. The GLARF rules implemented vary by language as follows. **English:** correcting/standardizing phrase boundaries and part of speech (POS); recognizing multiword expressions; marking subconstituents; labeling relations; incorporating NEs; regularizing infinitival, passives, relatives, VP deletion, predicative and numerous other constructions. **Chinese:** correcting/standardizing phrase boundaries and POS, marking subconstituents, labeling relations; regularizing copula constructions; incorporating NEs; recognizing dates and number expressions. **Japanese:** converting to PTB format; correcting/standardizing phrase boundaries and POS; labeling relations; processing NEs, double quote constructions, number phrases, common idioms, light verbs and copula constructions.

6 Discussion

Naturally, the treebank-based system outperformed parse-based system. The Charniak parser for English was trained on the Wall Street Journal corpus and can achieve about 90% accuracy on similar corpora, but lower accuracy on other genres. Differences between treebank and parser results for English were higher for LET and NAR genres than for the TEL because the system is not currently designed to handle TEL-specific features like disfluencies. All processors were trained on or initially designed for news corpora. Thus corpora out of this domain usually produce lower results. LET was easier as it consisted mainly of short simple sentences. In (Meyers et al., 2009), we evaluated our results on 40 Japanese sentences from the JENAAD corpus (Utiyama and Isahara, 2003) and achieved a higher F-score (90.6%) relative to the Kyoto corpus, as JENAAD tends to have fewer long complex sentences.

By using our answer key for multiple inputs, we discovered errors and consequently improved the quality of the answer keys. However, at times we were also compelled to *fork* the answer keys—given multiple correct answers, we needed to allow different answer keys corresponding to different inputs. For English, these items represent approximately 2% of the answer keys (there were a total

ID	Treebank				Parser			
	% Prec	% Rec	F	F-T	% Prec	% Rec	F	F-T
WSJ	$\frac{1238}{1491} = 83.0$	$\frac{1238}{1471} = 84.2$	83.6	87.1	$\frac{1164}{1452} = 80.2$	$\frac{1164}{1475} = 78.9$	79.5	81.8
LET	$\frac{419}{451} = 92.9$	$\frac{419}{454} = 92.3$	92.6	93.3	$\frac{390}{434} = 89.9$	$\frac{390}{454} = 85.9$	87.8	87.8
TEL	$\frac{478}{627} = 76.2$	$\frac{478}{589} = 81.2$	78.6	82.2	$\frac{439}{587} = 74.8$	$\frac{439}{589} = 74.5$	74.7	77.4
NAR	$\frac{817}{1013} = 80.7$	$\frac{817}{973} = 84.0$	82.3	84.1	$\frac{724}{957} = 75.7$	$\frac{724}{969} = 74.7$	75.2	76.1
CTB	$\frac{351}{400} = 87.8$	$\frac{351}{394} = 89.1$	88.4	88.7	$\frac{352}{403} = 87.3$	$\frac{352}{438} = 80.4$	83.7	83.7
KYO	$\frac{525}{575} = 91.3$	$\frac{525}{577} = 91.0$	91.1	91.1	$\frac{493}{581} = 84.9$	$\frac{493}{572} = 86.2$	85.5	87.8

Figure 2: Logic1 Scores

Ambiguity	Corp	Treebank	Parser
1. Tokenization	NAR	2+- + hour, 2+- + cent	2-hour, 2-cent
2. Tokenization	NAR	can't = can + n't	can't = ca + n't
3. Prefix?	KYO	大/big + 枠 /framework	大枠/the big picture
4. Encoding of zero	CTB	二 0 0 0 年/year 2000	二 000 年/year 2000
5. Attachment (relative)	LET	thousands [of people] [who face obstacles]	thousands of [people [who face obstacles]]
6. Attachment (PP)	LET	give a gift [to Goodwill]	give [a gift [to Goodwill]]
7. Conj Scope	TEL	[pearls or [beads of some sort of necklace]]	[[pearls or beads] of some sort of necklace]
8. Mod ambiguity	KYO	Relative Clause businesses that are varied	Adjectival Modifier various businesses
9. 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

Figure 3: Examples of Answer Key Divergences

of 74 4-tuples out of a total of 3487). Figure 3 lists examples of answer key divergences that we have found: (1) alternative tokenizations; (2) spurious differences in attachment and conjunction scope; and (3) ambiguities specific to our framework.

Examples 1 and 2 reflect different treatments of hyphenation and contractions in treebank specifications over time. Parsers trained on different treebanks will either keep hyphenated words together or separate more words at hyphens. The Treebank treatment of *can't* regularizes so that (*can* need not be differentiated from *ca*), whereas the parser treatment makes maintaining character offsets easier. In example 3, the Japanese parser recognizes a single word whereas the treebank divides it into a prefix plus stem. Example 4 is a case of differences in character encoding (zero).

Example 5 is a common case of spurious attachment ambiguity for English, where a transparent noun takes an *of* PP complement—nouns such as *form*, *variety* and *thousands* bear the feature *transparent* in the NOMLEX-PLUS dictionary (a NomBank dictionary based on NOMLEX (Macleod et al., 1998b)). The relative clause attaches either to the noun *thousands* or *people* and, therefore,

the subject gap of the relative is filled by either *thousands* or *people*. This ambiguity is spurious since there is no meaningful distinction between these two attachments. Example 6 is a case of attachment ambiguity due to a support construction (Meyers et al., 2004). The recipient of the gift will be *Goodwill* regardless of whether the PP is attached to *give* or *gift*. Thus there is not much sense in marking one attachment more correct than the other. Example 7 is a case of conjunction ambiguity—the context does not make it clear whether or not the pearls are part of a necklace or just the beads are. The distinction is of little consequence to the understanding of the narrative.

Example 8 is a case in which our grammar handles a case ambiguously: the prenominal adjective can be analyzed either as a simple noun plus adjective phrase meaning *various businesses* or as a noun plus relative clause meaning *businesses that are varied*. Example 9 is a common case in Chinese where the verb/noun distinction, while unclear, is not crucial to the meaning of the phrase – under either interpretation, 5 billion was exported.

7 Concluding Remarks

We have discussed challenges of automatic annotation when transducers of other annotation schemata are used as input. Models underlying different transducers approximate the original annotation in different ways, as do transducers trained on different corpora. We have found it necessary to allow for multiple *correct* answers, due to such differences, as well as, genuine and spurious ambiguities. In the future, we intend to investigate automatic ways of identifying and handling spurious ambiguities which are predictable, including examples like 5,6 and 7 in figure 3 involving transparent functors.

References

- E. Charniak. 2001. Immediate-head parsing for language models. In *ACL 2001*, pages 116–123.
- M. Collins. 1999. *Head-Driven Statistical Models for Natural Language Parsing*. Ph.D. thesis, University of Pennsylvania.
- D. Gildea and D. Jurafsky. 2002. Automatic Labeling of Semantic Roles. *Computational Linguistics*, 28:245–288.
- R. Grishman, D. Westbrook, and A. Meyers. 2005. Nyu’s english ace 2005 system description. In *ACE 2005 Evaluation Workshop*.
- J. Hajič, M. Ciaramita, R. Johansson, D. Kawahara, M. A. Martí, L. Màrquez, A. Meyers, J. Nivre, S. Padó, J. Štěpánek, P. Straňák, M. Surdeanu, N. Xue, and Y. Zhang. 2009. The CoNLL-2009 shared task: Syntactic and semantic dependencies in multiple languages. In *CoNLL-2009*, Boulder, Colorado, USA.
- Z. Huang and M. Harper. 2009. Self-training PCFG Grammars with Latent Annotations across Languages. In *EMNLP 2009*.
- H. Ji and R. Grishman. 2006. Analysis and Repair of Name Tagger Errors. In *COLING/ACL 2006*, Sydney, Australia.
- S. Kurohashi and M. Nagao. 1998. Building a Japanese parsed corpus while improving the parsing system. In *Proceedings of The 1st International Conference on Language Resources & Evaluation*, pages 719–724.
- C. Macleod, R. Grishman, and A. Meyers. 1998a. COMLEX Syntax. *Computers and the Humanities*, 31:459–481.
- C. Macleod, R. Grishman, A. Meyers, L. Barrett, and R. Reeves. 1998b. Nomlex: A lexicon of nominalizations. In *Proceedings of Euralex98*.
- A. Meyers, M. Kosaka, S. Sekine, R. Grishman, and S. Zhao. 2001. Parsing and GLARFing. In *Proceedings of RANLP-2001*, Tzigov Chark, Bulgaria.
- A. Meyers, R. Reeves, and Catherine Macleod. 2004. NP-External Arguments: A Study of Argument Sharing in English. In *The ACL 2004 Workshop on Multiword Expressions: Integrating Processing*, Barcelona, Spain.
- A. Meyers, M. Kosaka, N. Xue, H. Ji, A. Sun, S. Liao, and W. Xu. 2009. Automatic Recognition of Logical Relations for English, Chinese and Japanese in the GLARF Framework. In *SEW-2009 at NAACL-HLT-2009*.
- M. Surdeanu, R. Johansson, A. Meyers, L. Màrquez, and J. Nivre. 2008. The CoNLL-2008 Shared Task on Joint Parsing of Syntactic and Semantic Dependencies. In *Proceedings of the CoNLL-2008 Shared Task*, Manchester, GB.
- M. Utiyama and H. Isahara. 2003. Reliable Measures for Aligning Japanese-English News Articles and Sentences. In *ACL-2003*, pages 72–79.
- N. Xue, F. Xia, F. Chiou, and M. Palmer. 2005. The Penn Chinese Treebank: Phrase Structure Annotation of a Large Corpus. *Natural Language Engineering*.
- N. Xue. 2008. Labeling Chinese Predicates with Semantic roles. *Computational Linguistics*, 34:225–255.