

Automatic Recognition of Logical Relations for English, Chinese and Japanese

Adam Meyers[†], Michiko Kosaka[‡], Nianwen Xue[◇], Heng Ji^{*}, Ang Sun[†], Shasha Liao[†] and Wei Xu[†]
[†] New York University, [‡]Monmouth University, [◇]Brandeis University, ^{*} City University of New York

Abstract

We present a framework for representing three linguistic levels and systems for generating this representation. We focus on a logical level, like LFG’s F-structure, but compatible with Penn Treebanks. While less fine-grained than typical semantic role labeling approaches, our logical structure has several advantages: (1) it includes all words in all sentences, regardless of part of speech or semantic domain; and (2) it is easier to produce accurately. Our systems achieve 90% for English/Japanese News and 74.5% for Chinese News – these F-scores are nearly the same as those achieved for treebank-based parsing.

1 Introduction

For decades, computational linguists have paired a surface syntactic analysis with an analysis representing something “deeper”. The work of Harris (1968), Chomsky (1957) and many others showed that one could use these deeper analyses to regularize differences between ways of expressing the same idea. For statistical methods, these regularizations, in effect, reduce the number of significant differences between observable patterns in data and raise the frequency of each difference. Patterns are thus easier to learn from training data and easier to recognize in test data, thus somewhat compensating for the sparseness of data. In addition, deeper analyses are often considered semantic in nature because conceptually, two expressions that share the same regularized form also share some aspects of meaning. The specific details of this “deep” analysis have varied quite a bit, perhaps more than surface syntax.

In the 1970s and 1980s, Lexical Function Grammar’s (LFG) way of dividing C-structure (surface) and F-structure (deep) led to parsers such as (Hobbs and Grishman, 1976) which produced these two levels, typically in two stages. However, enthusiasm for these two-stage parsers was eclipsed by the advent of one stage parsers with much higher accuracy (about 90% vs about 60%), the now-popular treebank-based parsers including (Charniak, 2001; Collins, 1999) and many others. Currently, many different “deeper” levels are being manually annotated and automatically transduced, typically using surface parsing and other processors as input. One of the most popular, semantic role labels (annotation and transducers based on the annotation) characterize relations anchored by select predicate types like verbs (Palmer et al., 2005), nouns (Meyers et al., 2004a), discourse connectives (Miltasakaki et al., 2004) or those predicates that are part of particular semantic frames (Baker et al., 1998). The CONLL tasks for 2008 and 2009 (Surdeanu et al., 2008; Hajič et al., 2009) has focused on unifying many of these individual efforts to produce a logical structure for multiple parts of speech and multiple languages.

Like the CONLL shared task, we link surface levels to logical levels for multiple languages. However, there are several differences: (1) The logical structures produced automatically by our system can be expected to be more accurate than the comparable CONLL systems because our task involves predicting semantic roles with less fine-grained distinctions. Our English and Japanese results were higher than the CONLL 2009 SRL systems. Our English F-scores range from 76.3% (spoken) to 89.9% (News):

the best CONLL 2009 English scores were 73.31% (Brown) and 85.63% (WSJ). Our Japanese system scored 90.6%: the best CONLL 2009 Japanese score was 78.35%. Our Chinese system 74.5% scored 74.5%, 4 points lower than the best CONLL 2009 system (78.6%), probably due to our system’s failings, rather than the complexity of the task; (2) Each of the languages in our system uses the same linguistic framework, using the same types of relations, same analyses of comparable constructions, etc. In one case, this required a conversion from a different framework to our own. In contrast, the 2009 CONLL task puts several different frameworks into one compatible input format. (3) The logical structures produced by our system typically connect all the words in the sentence. While this is true for some of the CONLL 2009 languages, e.g., Czech, it is not true about all the languages. In particular, the CONLL 2009 English and Chinese logical structures only include noun and verb predicates.

In this paper, we will describe the GLARF framework (Grammatical and Logical Representation Framework) and a system for producing GLARF output (Meyers et al., 2001; Meyers, 2008). GLARF provides a logical structure for English, Chinese and Japanese with an F-score that is within a few percentage points of the best parsing results for that language. Like LFG’s (LFG) F-structure, our logical structure is less fine-grained than many of the popular semantic role labeling schemes, but also has two main advantages over these schemes: it is more reliable and it is more comprehensive in the sense that it covers all parts of speech and the resulting logical structure is a connected graph. Our approach has proved adequate for three genetically unrelated natural languages: English, Chinese and Japanese. It is thus a good candidate for additional languages with accurate parsers.

2 The GLARF framework

Our system creates a multi-tiered representation in the GLARF framework, combining the theory underlying the Penn Treebank for English (Marcus et al., 1994) and Chinese (Xue et al., 2005) (Chomskian linguistics of the 1970s and 1980s) with: (2) Relational Grammar’s graph-based way of representing “levels” as sequences of relations; (2) Fea-

ture structures in the style of Head-Driven Phrase Structure Grammar; and (3) The Z. Harris style goal of attempting to regularize multiple ways of saying the same thing into a single representation. Our approach differs from LFG F-structure in several ways: we have more than two levels; we have a different set of relational labels; and finally, our approach is designed to be compatible with the Penn Treebank framework and therefore, Penn-Treebank-based parsers. In addition, the expansion of our theory is governed more by available resources than by the underlying theory. As our main goal is to use our system to regularize data, we freely incorporate any analysis that fits this goal. Over time, we have found ways of incorporating Named Entities, PropBank, NomBank and the Penn Discourse Treebank. Our agenda also includes incorporating the results of other research efforts (Pustejovsky et al., 2005).

For each sentence, we generate a feature structure (FS) representing our most complete analysis. We distill a subset of this information into a dependency structure governed by theoretical assumptions, e.g., about identifying *functors* of phrases. Each GLARF dependency is between a functor and an argument, where the functor is the head of a phrase, conjunction, complementizer, or other function word. We have built applications that use each of these two representations, e.g., the dependency representation is used in (Shinyama, 2007) and the FS representation is used in (K. Parton and K. R. McKeown and R. Coyne and M. Diab and R. Grishman and D. Hakkani-Tür and M. Harper and H. Ji and W. Y. Ma and A. Meyers and S. Stolbach and A. Sun and G. Tür and W. Xu and S. Yarman, 2009).

In the dependency representation, each sentence is a set of 23 tuples, each 23-tuple characterizing up to three relations between two words: (1) a SURFACE relation, the relation between a functor and an argument in the parse of a sentence; (2) a LOGIC1 relation which regularizes for lexical and syntactic phenomena like passive, relative clauses, deleted subjects; and (3) a LOGIC2 relation corresponding to relations in PropBank, NomBank, and the Penn Discourse Treebank (PDTB). While the full output has all this information, we will limit this paper to a discussion of the LOGIC1 relations. Figure 1 is a 5 tuple subset of the 23 tuple GLARF analysis of the sentence *Who was eaten by Grendel?* (The full

L1	Surf	L2	Func	Arg
NIL	SENT	NIL	Who	was
PRD	PRD	NIL	was	eaten
COMP	COMP	ARG0	eaten	by
OBJ	NIL	ARG1	eaten	Who
NIL	OBJ	NIL	by	Grendel
SBJ	NIL	NIL	eaten	Grendel

Figure 1: 5-tuples: *Who was eaten by Grendel*

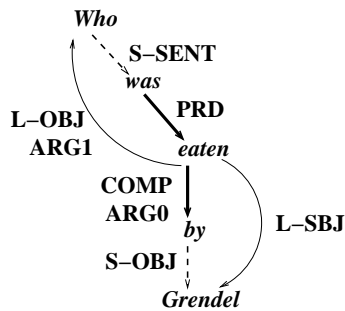


Figure 2: Graph of *Who was eaten by Grendel*

23 tuples include unique ids and fine-grained linguistic features). The fields listed are: logic1 label (L1), surface label (Surf), logic2 label (L2), functor (Func) and argument (Arg). NIL indicates that there is no relation of that type. Figure 2 represents this as a graph. For edges with two labels, the ARG0 or ARG1 label indicates a LOGIC2 relation. Edges with an L- prefix are LOGIC1 labels (the edges are curved); edges with S- prefixes are SURFACE relations (the edges are dashed); and other (thick) edges bear unprefix labels representing combined SURFACE/LOGIC1 relations. Deleting the dashed edges yields a LOGIC1 representation; deleting the curved edges yields a SURFACE representation; and a LOGIC2 consists of the edges labeled ARG0 and ARG1 relations, plus the surface subtrees rooted where the LOGIC2 edges terminate. Taken together, a sentence's SURFACE relations form a tree; the LOGIC1 relations form a directed acyclic graph; and the LOGIC2 relations form directed graphs with some cycles and, due to PDTB relations, may connect sentences to previous ones, e.g., adverbs like *however*, take the previous sentence as one of their arguments.

LOGIC1 relations (based on Relational Grammar) regularize across grammatical and lexical al-

ternations. For example, subcategorized verbal arguments include: SBJect, OBJect and IND-OBJ (indirect Object), COMPLEMENT, PRT (Particle), PRD (predicative complement). Other verbal modifiers include AUXilliary, PARENthetical, ADVerbial. In contrast, FrameNet and PropBank make finer distinctions. Both PP arguments of *consulted* in *John consulted with Mary about the project* bear COMP relations with the verb in GLARF, but would have distinct labels in both PropBank and FrameNet. Thus Semantic Role Labeling (SRL) should be more difficult than recognizing LOGIC1 relations.

Beginning with Penn Treebank II, Penn Treebank annotation includes Function tags, hyphenated additions to phrasal categories which indicate their function. There are several types of function tags:

- **Argument Tags** such as SBJ, OBJ, IO (IND-OBJ), CLR (COMP) and PRD—These are limited to verbal relations and not all are used in all treebanks. For example, OBJ and IO are used in the Chinese, but not the English treebank. These labels can often be directly translated into GLARF LOGIC1 relations.
- **Adjunct Tags** such as ADV, TMP, DIR, LOC, MNR, PRP—These tags often translate into a single LOGIC1 tag (ADV). However, some of these also correspond to LOGIC1 arguments. In particular, some DIR and MNR tags are realized as LOGIC1 COMP relations (based on dictionary entries). The fine grained semantic distinctions are maintained in other features that are part of the GLARF description.

In addition, GLARF treats Penn's PRN phrasal category as a relation rather than a phrasal category. For example, given a sentence like, *Banana ketchup, the agency claims, is very nutritious*, the phrase *the agency claims* is analyzed as an S(entence) in GLARF bearing a (surface) PAREN relation to the main clause. Furthermore, the whole sentence is a COMP of the verb *claims*. Since PAREN is a SURFACE relation, not a LOGIC1 relation, there is no LOGIC1 cycle as shown by the set of 5-tuples in Figure 3— a cycle only exists if you include both SURFACE and LOGIC1 relations in a single graph.

Another important feature of the GLARF framework is *transparency*, a term originating from N.

L1	Surf	L2	Func	Arg
NIL	SBJ	ARG1	is	ketchup
PRD	PRD	ARG2	is	nutritious
SBJ	NIL	NIL	nutritious	Ketchup
ADV	ADV	NIL	nutritious	very
N-POS	N-POS	NIL	ketchup	Banana
NIL	PAREN	NIL	is	claims
SBJ	SBJ	ARG0	claims	agency
Q-POS	Q-POS	NIL	agency	the
COMP	NIL	ARG1	claims	is

Figure 3: 5-tuples: *Banana Ketchup, the agency claims, is very nutritious*

L1	Surf	L2	Func	Arg
SBJ	SBJ	ARG0	ate	and
OBJ	OBJ	ARG1	ate	box
CONJ	CONJ	NIL	and	John
CONJ	CONJ	NIL	and	Mary
COMP	COMP	NIL	box	of
Q-POS	Q-POS	NIL	box	the
OBJ	OBJ	NIL	of	cookies

Figure 4: 5-tuples: *John and Mary ate the box of cookies*

Sager’s unpublished work. A relation between two words is transparent if: the functor fails to characterize the selectional properties of the phrase (or sub-graph in a Dependency Analysis), but its argument does. For example, relations between conjunctions (e.g., *and, or, but*) and their conjuncts are transparent CONJ relations. Thus although *and* links together *John* and *Mary*, it is these dependents that determine that the resulting phrase is noun-like (an NP in phrase structure terminology) and sentient (and thus can occur as the subject of verbs like *ate*). Another common example of transparent relations are the relations connecting certain nouns and the prepositional objects under them, e.g., *the box of cookies* is edible, because cookies are edible even though boxes are not. These features are marked in the NOMLEX-PLUS dictionary (Meyers et al., 2004b). In Figure 4, we represent transparent relations, by prefixing the LOGIC1 label with asterisks.

The above description most accurately describes English GLARF. However, Chinese GLARF has most of the same properties, the main exception being that PDTB arguments are not currently marked.

For Japanese, we have only a preliminary representation of LOGIC2 relations and they are not derived from PropBank/NomBank/PDTB.

2.1 Scoring the LOGIC1 Structure

For purposes of scoring, we chose to focus on LOGIC1 relations, our proposed high-performance level of semantics. We scored with respect to: the LOGIC1 relational label, the identity of the functor and the argument, and whether the relation is transparent or not. If the system output differs in any of these respects, the relation is marked wrong. The following sections will briefly describe each system and present an evaluation of its results.

The answering keys for each language were created by native speakers editing system output, as represented similarly to the examples in this paper, although part of speech is included for added clarity. In addition, as we attempted to evaluate logical relation (or dependency) accuracy independent of sentence splitting. We obtained sentence divisions from data providers and treebank annotation for all the Japanese and most of the English data, but used automatic sentence divisions for the English BLOG data. For the Chinese, we omitted several sentences from our evaluation set due to incorrect sentence splits. The English and Japanese answer keys were annotated by single native speakers expert in GLARF. The Chinese data was annotated by several native speakers and may have been subject to some interannotator agreement difficulties (which we plan to investigate further for the final version of this paper). Currently, correcting system output is the best way to create answer keys due to certain ambiguities in the framework, some of which we hope to incorporate into future scoring procedures. For example, consider the interpretation of the phrase *five acres of land in England* with respect to PP attachment. The difference in meaning between attaching the PP *in England* to *acres* or to *land* is too subtle for these authors—we have difficulty imagining situations where one statement would be accurate and the other would not. This ambiguity is completely predictable because *acres* is a transparent noun and similar ambiguities hold for all such cases where a transparent noun takes a complement and is followed by a PP attachment. We believe that a more complex scoring program could account for

most of these cases. Similar complexities arise for coordination and several other phenomena.

3 English GLARF

We generate English GLARF output by applying a procedure that combines:

1. The output of the 2005 version of the Charniak parser described in (Charniak, 2001), which label precision and recall scores in the 85% range. The updated version of the parser seems to perform closer to 90% on News data and perform lower on other genres. That performance would reflect reports on other versions of the Charniak parser for which statistics are available (Foster and van Genabith, 2008).
2. Named entity (NE) tags from the JET NE system (Ji and Grishman, 2006), which achieves F-scores ranging 86%-91% on newswire for both English and Chinese (depending on Epoch). The JET system identifies seven classes of NEs: Person, GPE, Location, Organization, Facility, Weapon and Vehicle.
3. Machine Readable dictionaries: COMLEX (Macleod et al., 1998), NOMBANK dictionaries (from <http://nlp.cs.nyu.edu/meyers/nombank/>) and others.
4. A sequence of hand-written rules (citations omitted) such that: (1) the first set of rules convert the Penn Treebank into a Feature Structure representation; and (2) each rule N after the first rule is applied to an entire Feature Structure that is the output of rule $N - 1$.

For this paper, we evaluated the English output for several different genres, all of which approximately track parsing results for that genre. For written genres, we chose between 40 and 50 sentences. For speech transcripts, we chose 100 sentences—we chose this larger number because a lot of so-called sentences contained text with empty logical descriptions, e.g., single word utterances contain no relations between pairs of words. Each text comes from a different genre. For NEWS text, we used 50 sentences from the aligned Japanese-English data created as part of the JENAAD corpus (Utiyama

Genre	Prec	Rec	F
NEWS	$\frac{731}{815} = 89.7\%$	$\frac{715}{812} = 90.0\%$	89.9%
BLOG	$\frac{704}{844} = 83.4\%$	$\frac{704}{899} = 78.3\%$	80.8%
LETT	$\frac{392}{434} = 90.3\%$	$\frac{392}{449} = 87.3\%$	88.8%
TELE	$\frac{472}{604} = 78.1\%$	$\frac{472}{610} = 77.4\%$	77.8%
NARR	$\frac{732}{959} = 76.3\%$	$\frac{732}{964} = 75.9\%$	76.1%

Table 1: English Aggregate Scores

Corpus	Prec	Rec	F	Sents
NEWS	90.5%	90.8%	90.6%	50
BLOG	84.1%	79.6%	81.7%	46
LETT	93.9%	89.2%	91.4%	46
TELE	81.4%	83.2%	84.9%	103
NARR	77.1%	78.1%	79.5%	100

Table 2: English Score per Sentence

and Isahara, 2003); the web text (BLOGs) was taken from some corpora provided by the Linguistic Data Consortium through the GALE (<http://projects.ldc.upenn.edu/gale/>) program; the LETTer genre (a letter from Good Will) was taken from the ICIC Corpus of Fundraising Texts (Indiana Center for Intercultural Communication); Finally, we chose two spoken language transcripts: a TELEphone conversation from the Switchboard Corpus (http://www.ldc.upenn.edu/Catalog/readme_files/switchboard.readme.html) and one NAR-Rative from the Charlotte Narrative and Conversation Collection (<http://newsouthvoices.uncc.edu/cncc.php>). In both cases, we assumed perfect sentence splitting (based on Penn Treebank annotation). The ICIC, Switchboard and Charlotte texts that we used are part of the Open American National Corpus (OANC), in particular, the SIGANN shared subcorpus of the OANC (<http://nlp.cs.nyu.edu/wiki/corpuswg/ULA-OANC-1>) (Meyers et al., 2007).

Comparable work for English includes: (1) (Gabbard et al., 2006), a system which reproduces the function tags of the Penn Treebank with 89% accuracy and empty categories (and their antecedents) with varying accuracies ranging from 82.2% to 96.3%, excluding null complementizers, as these are theory-internal and have no value for filling gaps. (2) Current systems that generate LFG F-structure

such as (Wagner et al., 2007) which achieve an F score of 91.1 on the F-structure PRED relations, which are similar to our LOGIC1 relations.

4 Chinese GLARF

The Chinese GLARF program takes a Chinese Treebank-style syntactic parse and the output of a Chinese PropBanker (Xue, 2008) as input, and attempts to determine the relations between the head and its dependents within each constituent. It does this by first exploiting the structural information and detecting six broad categories of syntactic relations that hold between the head and its dependents. These are *predication*, *modification*, *complementation*, *coordination*, *auxiliary*, and *flat*. Predication holds at the clause level between the subject and the predicate, where the predicate is considered to be the head and the subject is considered to be the dependent. Modification can also hold mainly within NPs and VPs, where the dependents are modifiers of the NP head or adjuncts to the head verb. Coordination holds almost for all phrasal categories where each non-punctuation child within this constituent is either conjunction or a conjunct. The head in a coordination structure is underspecified and can be either a conjunct or a conjunction depending on the grammatical framework. Complementation holds between a head and its complement, with the complement usually being a core argument of the head. For example, inside a PP, the preposition is the head and the phrase or clause it takes is the dependent. An auxiliary structure is one where the auxiliary takes a VP as its complement. This structure is identified so that the auxiliary and the verb it modifies can form a verb group in the GLARF framework. Flat structures are structures where a constituent has no meaningful internal structure, which is possible in a small number of cases. After these six broad categories of relations are identified, more fine-grained relation can be detected with additional information. Figure 5 is a sample 4-tuple for a Chinese translation of the sentence in figure 3.

For the results reported in Table 3, we used the Harper and Huang parser described in (Harper and Huang, Forthcoming) which can achieve F-scores as high as 85.2%, in combination with information about named entities from the output of the

L1	Surf	L2	Func	Arg
SBJ	SBJ	ARG0	说 claims	代理 agency
COMP	COMP	NIL	说 claims	是 is
NIL	SBJ	ARG1	是 is	酱 ketchup
N-POS	N-POS	NIL	酱 ketchup	香蕉 banana
PRD	PRD	ARG2	是 is	有 have
ADV	ADV	NIL	有 have	很 very
OBJ	OBJ	ARG2	有 have	营养 nutrition
SBJ	NIL	ARG1	有 have	酱 ketchup
SENT	NIL	NIL	的 DE	有 have

代理说, 香蕉酱是很有营养的。

Figure 5: Agency claims, Banana Ketchup is very have nutrition DE.

JET Named Entity tagger for Chinese (86%-91% F-measure as per section 3). We used the NE tags to adjust the parts of speech and the phrasal boundaries of named entities (we do the same with English). As shown in Table 3, we tried two versions of the Harper and Huang parser, one which adds function tags to the output and one that does not. The Chinese GLARF system scores significantly (13.9% F-score) higher given function tagged input, than parser output without function tags. Our current score is about 10 points lower than the parser score. Our initial error analysis suggests that the most common forms of errors involve: (1) the processing of long NPs; (2) segmentation and POS errors; (3) conjunction scope; and (4) modifier attachment.

5 Japanese GLARF

For Japanese, we process text with the KNP parser (Kurohashi and Nagao, 1998) and convert the output into the GLARF framework. The KNP/Kyoto Corpus framework is a Japanese-specific Dependency framework, very different from the Penn Treebank framework used for the other systems. Processing in Japanese proceeds as follows: (1) we process the Japanese with the Juman segmenter (Kuro-

Type	Prec	Rec	F
No Function Tags Version			
Aggr	$\frac{843}{1374} = 61.4\%$	$\frac{843}{1352} = 62.4\%$	61.8%
Aver	62.3%	63.5%	63.6%
Function Tags Version			
Aggr	$\frac{1031}{1415} = 72.9\%$	$\frac{1031}{1352} = 76.3\%$	74.5%
Aver	73.0%	75.3%	74.9%

Table 3: 53 Chinese Newswire Sentences: Aggregate and Average Sentence Scores

hashi et al., 1994) and KNP parser 2.0 (Kurohashi and Nagao, 1998), which has reported accuracy of 91.32% F score for dependency accuracy, as reported in (Noro et al., 2005). As is standard in Japanese linguistics, the KNP/Kyoto Corpus (K) framework uses a dependency analysis that has some features of a phrase structure analysis. In particular, the dependency relations are between *bunsetsu*, small constituents which include a head word and some number of modifiers which are typically function words (particles, auxiliaries, etc.), but can also be prenominal noun modifiers. *Bunsetsu* can also include multiple words in the case of names. The K framework differentiates types of dependencies into: the normal head-argument variety, coordination (or parallel) and apposition. We convert the head-argument variety of dependency straightforwardly into a phrase consisting of the head and all the arguments. In a similar way, appositive relations could be represented using an APPOSITIVE relation (as is currently done with English). In the case of *bunsetsu*, the task is to choose a head and label the other constituents—This is very similar to our task of labeling and subdividing the flat noun phrases of the English Penn Treebank. Conjunction is a little different because the K analysis assumes that the final conjunct is the functor, rather than a conjunction. We automatically changed this analysis to be the same as it is for English and Chinese. When there was no actual conjunction, we created a theory-internal NULL conjunction. The final stages include: (1) processing conjunction and apposition, including recognizing cases that the parser does not recognize; (2) correcting parts of speech; (3) labeling all relations between arguments and heads; (4) recognizing and labeling special constituent types

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

生命、財産を守ることは、国家の責務だ。

Figure 6: It is the state’s duty to protect lives and assets.

Type	Prec	Rec	F
Aggr	$\frac{764}{843} = 91.0\%$	$\frac{764}{840} = 90.6\%$	90.8%
Aver	90.7%	90.6%	90.6%

Table 4: 40 Japanese Sentences from JENAA Corpus: Aggregate and Average Sentence Scores

such as Named Entities, double quote constituents and number phrases (*twenty one*); (5) handling common idioms; and (6) processing light verb and copula constructions.

Figure 6 is a sample 4-tuple for a Japanese sentence meaning *It is the state’s duty to protect lives and assets*. Conjunction is handled as discussed above, using an invisible NULL conjunction and transparent (asterisked) logical CONJ relations. Copulas in all three languages take surface subjects, which are the LOGIC1 subjects of the PRD argument of the copula. We have left out glosses for the particles, which act solely as case markers and help us identify the grammatical relation.

We scored Japanese GLARF on forty sentences of the Japanese side of the JENAA data (25 of which are parallel with the English sentences scored). Like the English, the F score is very close to the parsing scores achieved by the parser.

6 Concluding Remarks and Future Work

In this paper, we have described three systems for generating GLARF representations automatically from text, each system combines the output of a parser and possibly some other processor (segmenter, Named Entity Recognizer, PropBanker, etc.) and creates a logical representation of the sentence. Dictionaries, word lists, and various other resources are used, in conjunction with hand written rules. In each case, the results are very close to parsing accuracy. These logical structures are in the same annotation framework, using the same labeling scheme and the same analysis for key types of constructions. There are several advantages to our approach over other characterizations of logical structure: (1) our representation is among the most accurate and reliable; (2) our representation connects all the words in the sentence; and (3) having the same representation for multiple languages facilitates running the same procedures in multiple languages and creating multilingual applications.

The English system was developed for the News genre, specifically the Penn Treebank Wall Street Journal Corpus. We are therefore considering adding rules to better handle constructions that appear in other genres, but not news. The experiments describe here should go a long way towards achieving this goal. We are also considering experiments with parsers tailored to particular genres and/or parsers that add function tags (Harper et al., 2005). In addition, our current GLARF system uses internal Propbank/NomBank rules, which have good precision, but low recall. We expect that we achieve better results if we incorporate the output of state of the art SRL systems, although we would have to conduct experiments as to whether or not we can improve such results with additional rules.

We developed the English system over the course of eight years or so. In contrast, the Chinese and Japanese systems are newer and considerably less time was spent developing them. Thus they currently do not represent as many regularizations. One obstacle is that we do not currently use subcategorization dictionaries for either language, while we have several for English. In particular, these would be helpful in predicting and filling relative clause and others gaps. We are considering auto-

matically acquiring simple dictionaries by recording frequently occurring argument types of verbs over a larger corpus, e.g., along the lines of (Kawahara and Kurohashi, 2002). In addition, existing Japanese dictionaries such as the IPAL (monolingual) dictionary (technology Promotion Agency, 1987) or previously acquired case information reported in (Kawahara and Kurohashi, 2002).

Finally, we are investigating several avenues for using this system output for Machine Translation (MT) including: (1) aiding word alignment for other MT system (Wang et al., 2007); and (2) aiding the creation various MT models involving analyzed text, e.g., (Gildea, 2004; Shen et al., 2008).

Acknowledgments

This work was supported by NSF Grant IIS-0534700 Structure Alignment-based MT.

References

- C. F. Baker, C. J. Fillmore, and J. B. Lowe. 1998. The Berkeley FrameNet Project. In *Coling-ACL98*, pages 86–90.
- E. Charniak. 2001. Immediate-head parsing for language models. In *ACL 2001*, pages 116–123.
- N. Chomsky. 1957. *Syntactic Structures*. Mouton, The Hague.
- M. Collins. 1999. *Head-Driven Statistical Models for Natural Language Parsing*. Ph.D. thesis, University of Pennsylvania.
- J. Foster and J. van Genabith. 2008. Parser Evaluation and the BNC: 4 Parsers and 3 Evaluation Metrics. In *LREC 2008*, Marrakech, Morocco.
- R. Gabbard, M. Marcus, and S. Kulick. 2006. Fully parsing the penn treebank. In *NAACL/HLT*, pages 184–191.
- D. Gildea. 2004. Dependencies vs. Constituents for Tree-Based Alignment. In *EMNLP*, Barcelona.
- 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.
- M. Harper and Z. Huang. Forthcoming. Chinese Statistical Parsing. In J. Olive, editor, *Global Autonomous Language Exploitation*. Publisher to be Announced.
- M. Harper, B. Dorr, J. Hale, B. Roark, I. Shafran, M. Lease, Y. Liu, M. Snover, L. Yung, A. Krasnyanskaya, and R. Stewart. 2005. Parsing and Spoken

- Structural Event. Technical Report, The John-Hopkins University, 2005 Summer Research Workshop.
- Z. Harris. 1968. *Mathematical Structures of Language*. Wiley-Interscience, New York.
- J. R. Hobbs and R. Grishman. 1976. The Automatic Transformational Analysis of English Sentences: An Implementation. *International Journal of Computer Mathematics*, 5:267–283.
- H. Ji and R. Grishman. 2006. Analysis and Repair of Name Tagger Errors. In *COLING/ACL 2006*, Sydney, Australia.
- K. Parton and K. R. McKeown and R. Coyne and M. Diab and R. Grishman and D. Hakkani-Tür and M. Harper and H. Ji and W. Y. Ma and A. Meyers and S. Stolbach and A. Sun and G. Tür and W. Xu and S. Yarman. 2009. Who, What, When, Where, Why? Comparing Multiple Approaches to the Cross-Lingual 5W Task. In *CLIAWS3*.
- D. Kawahara and S. Kurohashi. 2002. Fertilization of Case Frame Dictionary for Robust Japanese Case Analysis. In *Proc. of COLING 2002*.
- 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.
- S. Kurohashi, T. Nakamura, Y. Matsumoto, and M. Nagao. 1994. Improvements of Japanese Morphological Analyzer JUMAN. In *Proc. of International Workshop on Sharable Natural Language Resources (SNLR)*, pages 22–28.
- C. Macleod, R. Grishman, and A. Meyers. 1998. COMLEX Syntax. *Computers and the Humanities*, 31:459–481.
- M. P. Marcus, B. Santorini, and M. A. Marcinkiewicz. 1994. Building a large annotated corpus of english: The penn treebank. *Computational Linguistics*, 19(2):313–330.
- 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, C. Macleod, R. Szekely, V. Zielinska, B. Young, and R. Grishman. 2004a. The NomBank Project: An Interim Report. In *NAACL/HLT 2004 Workshop Frontiers in Corpus Annotation*, Boston.
- A. Meyers, R. Reeves, Catherine Macleod, Rachel Szekely, Veronkia Zielinska, and Brian Young. 2004b. The Cross-Breeding of Dictionaries. In *Proceedings of LREC-2004*, Lisbon, Portugal.
- A. Meyers, N. Ide, L. Denoyer, and Y. Shinyama. 2007. The shared corpora working group report. In *Proceedings of The Linguistic Annotation Workshop, ACL 2007*, pages 184–190, Prague, Czech Republic.
- A. Meyers. 2008. Using treebank, dictionaries and glarf to improve nombank annotation. In *Proceedings of The Linguistic Annotation Workshop, LREC 2008*, Marrakesh, Morocco.
- E. Miltsakaki, A. Joshi, R. Prasad, and B. Webber. 2004. Annotating discourse connectives and their arguments. In A. Meyers, editor, *NAACL/HLT 2004 Workshop: Frontiers in Corpus Annotation*, pages 9–16, Boston, Massachusetts, USA, May 2 - May 7. Association for Computational Linguistics.
- T. Noro, C. Koike, T. Hashimoto, T. Tokunaga, and Hozumi Tanaka. 2005. Evaluation of a Japanese CFG Derived from a Syntactically Annotated corpus with Respect to Dependency Measures. In *2005 Workshop on Treebanks and Linguistic theories*, pages 115–126.
- M. Palmer, D. Gildea, and P. Kingsbury. 2005. The Proposition Bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106.
- J. Pustejovsky, A. Meyers, M. Palmer, and M. Poesio. 2005. Merging PropBank, NomBank, TimeBank, Penn Discourse Treebank and Coreference. In *ACL 2005 Workshop: Frontiers in Corpus Annotation II: Pie in the Sky*.
- L. Shen, J. Xu, and R. Weischedel. 2008. A New String-to-Dependency Machine Translation Algorithm with a Target Dependency Language Model. In *ACL 2008*.
- Y. Shinyama. 2007. *Being Lazy and Preemptive at Learning toward Information Extraction*. Ph.D. thesis, NYU.
- M. Surdeanu, R. Johansson, A. Meyers, Ll. Márquez, 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.
- Information technology Promotion Agency. 1987. IPA Lexicon of the Japanese Language for Computers IPAL (Basic Verbs). (in Japanese).
- M. Utiyama and H. Isahara. 2003. Reliable Measures for Aligning Japanese-English News Articles and Sentences. In *ACL-2003*, pages 72–79.
- J. Wagner, D. Seddah, J. Foster, and J. van Genabith. 2007. C-Structures and F-Structures for the British National Corpus. In *Proceedings of the Twelfth International Lexical Functional Grammar Conference*, Stanford. CSLI Publications.
- C. Wang, M. Collins, and P. Koehn. 2007. Chinese syntactic reordering for statistical machine translation. In *EMNLP-CoNLL 2007*, pages 737–745.
- N. Xue, F. Xia, F. Chiou, and M. Palmer. 2005. The Penn Chinese TreeBank: Phrase Structure Annotation of a Large Corpus. *Natural Language Engineering*, 11:207–238.
- N. Xue. 2008. Labeling Chinese Predicates with Semantic roles. *Computational Linguistics*, 34:225–255.