Toward a cross-framework parser annotation standard

Dan Flickinger
CSLI, Stanford University
danf@stanford.edu

Abstract

Efficient and precise comparison of parser results across frameworks will require a negotiated agreement on a target representation which embodies a good balance of three competing dimensions: consistency, clarity, and flexibility. The various annotations provided in the COLING-08 shared task for the ten 'required' Wall Street Journal sentences can serve as a useful basis for these negotations. While there is of course substantial overlap in the content of the various schemes for these sentences, no one of the schemes is ideal. This paper presents some desiderata for a negotiated target annotation scheme for which straightforward mappings can be constructed from each of the supplied annotation schemes.

1 Introduction

Efficient and precise comparison of parser results across frameworks will require a negotiated agreement on a target representation which embodies a good balance of three competing dimensions: consistency, clarity, and flexibility. The various annotations provided in the COLING-08 shared task for the ten 'required' Wall Street Journal sentences can serve as a useful basis for these negotations. While there is of course substantial overlap in the content of the various schemes for these sentences, no one of the schemes is ideal, containing either too much or too little detail, or sometimes both.

2 Predicate-argument structures, not labelled bracketings

Competing linguistic frameworks can vary dramatically in the syntactic structures they assign to sentences, and this variation makes cross-framework comparison of labelled bracketings difficult and in the limit uninteresting. The syntactic structures of Combinatory Categorial Grammar (CCG: Steedman (2000), Hockenmaier (2003), Clark and Curran (2003)), for example, contrast sharply with those of the Penn Treebank Marcus et al. (1993), and the PTB structures differ in many less dramatic though equally important details from those assigned in Lexical Functional Grammar (LFG: Bresnan and Kaplan (1982)) or Head-driven Phrase Structure Grammar (HPSG: Pollard and Sag (1994)). We even find variation in the assignments of part-of-speech tags for individual tokens, for example with words like "missionary" or "classical" treated as adjectives in some of the annotations and as nouns in others. Furthermore, a simple labelled bracketing of surface tokens obscures the fact that a single syntactic constituent can fill multiple roles in the logical structure expressed by a sentence, as with controlled subjects, relative clauses, appositives, coordination, etc. More detailed discussions of the obstacles to directly comparing syntactic structures include Preiss (2003), Clark and Curran (2007), and most recently Sagae et al. (2008).

Since it is this underlying logical content that we seek when parsing a sentence, the target annotation for cross-framework comparison should not include marking of syntactic constituents, but focus instead on the predicate argument structures determined by the syntactic analysis, as proposed ten years ago by Carroll et al. (1998). Several of

Coling 2008: Proceedings of the workshop on Cross-Framework and Cross-Domain Parser Evaluation, pages 24–28
Manchester, August 2008

the annotations provided in the shared task already do this, providing a good set of starting points for negotiating a common target.

3 General annotation characteristics

Some of the issues in need of negotiation are quite general in nature, while many others involve specific phenomena. First, the general ones:

3.1 Unique identifiers

Since a given word can appear multiple times within a single sentence, each token-derived element of the annotation needs a unique identifier. Some of the supplied annotations use the token position in the sentence for this purpose, but this is not general enough to support competing hypotheses about the number of tokens in a sentence. A sharp example of this is the word *pixie-like* in sentence 56, which one of the annotations (CONLL08) analyzes as two tokens, quite reasonably, since *-like* is a fully productive compounding element. So a better candidate for the unique identifier for each annotation element would be the initial *character* position of the source token in the original sentence, including spaces and punctuation marks as characters. Thus in the sentence *the dog slept* the annotation elements would be *the-1*, *dog-5*, and *slept-9*. The original sentences in this shared task were presented with spaces added around punctuation, and before "n't". so the character positions for this task would be computed taking this input as given. Using character positions rather than token positions would also better accommodate differing treatments of multi-word expressions, as for example with *Los Angeles* in sentence 9, which most of the supplied schemes annotate as two tokens with *Los* modifying *Angeles*, but which PARC treats as a single entity.

3.2 One token in multiple roles

Most of the supplied annotations include some notational convention to record the fact that (a phrase headed by) a single token can fill more than one logical role at the predicate-argument level of representation. This is clear for controlled subjects as in the one for *play* in sentence 53: *"doesn't have to play...concertos"*, and equally clear for the missing objects in *tough*-type adjective phrases, like the object of *apply* in sentence 133: *"impossible to apply"*. This multiple filling of roles by a single syntactic constituent can be readily expressed in a target annotation of the predicate argument structure if the token heading that constituent bears the unique positional identifier which has already been motivated above. Supplied annotation schemes that already directly employ this approach include PARC and Stanford, and the necessary positional information is also readily available in the CCG-PA, HPSG-PA, and CONLL08 schemes, though not in the RASP-GR or PTB notations. It will be desirable to employ this same convention for the logical dependencies in other constructions with missing arguments, including relative clauses, other unbounded dependencies like questions, and comparative constructions like sentence 608's *than President Bush has allowed* __.

3.3 Stem vs surface form

Some of the supplied annotations (CCG-PA, RASP-GR, and Stanford) simply use the surface forms of the tokens as the elements of relations, while most of the others identify the stem forms for each token. While stemming might introduce an additional source of inconsistency in the annotations, the resulting annotations will be better normalized if the stems rather than the surface forms of words are used. This normalization would also open the door to making such annotations more suitable for validation by reasoning engines, or for later word-sense annotation, or for applications.

3.4 Identification of root

Most but not all of the supplied annotation schemes identify which token supplies the outermost predication for the sentence, either directly or indirectly. An explicit marking of this outermost element, typically the finite verb of the main clause of a sentence, should be included in the target annotation, since it avoids the spurious ambiguity found for example in the HPSG-PA annotation for sentence 22, which looks like it would be identical for both of the following two sentences:

- *Not all those who wrote oppose the changes* .
- *Not all those who oppose the changes wrote* .

3.5 Properties of entities and events

Some of the supplied annotation schemes include information about morphosyntactically marked properties of nouns and verbs, including person, number, gender, tense, and aspect. Providing for explicit marking of these properties in a common

target annotation is desirable, at least to the level of detail adopted by several of the supplied schemes.

While several of the supplied annotation schemes marked some morphosyntactic properties some of the time, the PARC annotation of positive degree for all adjectives reminds us that it would be useful to adopt a notion of default values for these properties in the target annotation. These defaults would be explicitly defined once, and then only non-default values would need to be marked explicitly in the annotation for a given sentence. For example, the PARC annotation marks the 'perf' (perfect) attribute for a verb only when it has a positive value, implicitly using the negative value as the default. This use of defaults would improve the readability of the target annotation without any loss of information.

Marking of the contrast between declarative, interrogative, and imperative clauses is included in some but not all of the annotation schemes. Since this contrast is highly salient and (almost always) easily determined, it should be marked explicitly in the target annotation, at least for the main clause.

3.6 Named entities

The supplied annotations represent a variety of approaches to the treatment of named entities where multiple tokens comprise the relevant noun phrase, as in sentence 53's *"The oboist Heinz Holliger"*. Several schemes treat both *oboist* and *Heinz* simply as modifiers of *Holliger*, drawing no distinction between the two. The PARC and PTB annotations identify *Heinz Holliger* as a named entity, with *oboist* as a modifier, and only the CONLL08 scheme analyses this expression as an apposition, with *oboist* as the head predicate of the whole PN. Since complex proper names appear frequently with modifiers and in apposition constructions, and since competing syntactic and semantic analyses can be argued for many such constituents, the target annotation should contain enough detail to illuminate the substantive differences without exaggerating them. Interestingly, this suggests that the evaluation of a given analysis in comparison with a gold standard in the target annotation may require some computation of near-equivalence at least for entities in complex noun phrases. If scheme A treats *Holliger* as the head token for use in external dependencies involving the above noun phrase, while scheme B treats *oboist* as the head token, it will be important in evaluation to exploit the fact that both schemes each establish some relation between *oboist* and *Holliger* which can be interpreted as substitutional equivalence with respect to those external dependencies. This means that even when a target annotation scheme has been agreed upon, and a mapping defined to convert a native annotated analyis into a target annotation, it will still be necessary to create non-trivial software which can evaluate the mapped analysis against a gold standard analysis.

4 Notational conventions to be negotiated

A number of notational conventions will have to be negotiated for a common target annotation scheme, ranging from quite general design decisions to details about very specific linguistic phenomena.

4.1 Naming of arguments and relations

It seems plausible that agreement could be reached quickly on the names for at least the core grammatical functions of subject, direct object, indirect object, and verbal complement, and perhaps also on the names for adjectival and adverbial modifiers. Prepositions are more challenging, since they are very often two-place relations, and often live on the blurry border between arguments and adjuncts. For example, most of the supplied annotation schemes treated the *by*-PP following *moved* in sentence 608 as a marker for the logical subject of the passive verb, but this was at least not clear in the CCG-PA annotation. In sentence 56, there was variation in how the *from* and *to* PPs were annotated, with CONLL08 making the two *to* PPs dependents of the *from* PP rather than of the verb *range*.

Some of the supplied annotation schemes introduced reasonable but idiosyncratic names for other frequently occurring relations or dependencies such as relative clauses, appositives, noun-noun compounds, and subordinate clauses. An inventory of these frequently occurring phenomena should be constructed, and a target name negotiated for each, recognizing that there will always be a long tail of less frequently occurring phenomena where names will not (yet) have been negotiated.

4.2 Coordination

Perhaps the single most frequent source of apparent incompatibility in the supplied annotations for the ten required sentences in this task involves coordination. Some schemes, like HPSG-PA and

Stanford, treat the first conjunct as the primary entity which participates in other predications, with the other conjunct(s) dependent on the first, though even here they usually (but not always) distribute conjoined verbal arguments with separate predications for each conjunct. Some schemes, like the PTB, PARC, and RASP-GR, represent the grouping of three or more conjuncts as flat, while others like the Stanford scheme represent them as pairs. Most schemes make each conjunction word itself explicit, but for example the PARC annotation of 866 marks only one occurrence of *and* even though this three-part coordinate structure includes two explicit conjunctions.

While the distribution of conjoined elements in coordinate structures may be the most practical target annotation, it should at least be noted that this approach will not accommodate collective readings of coordinate NPs as in well-known examples like *"Tom and Mary carried the piano upstairs."* But the alternative, to introduce a new conjoined entity for every coordinate structure, may be too abstract to find common support among developers of current annotation schemes, and perhaps not worth the effort at present.

However, it should be possible to come to agreement on how to annotate the distribution of conjoined elements consistently, such that it is clear both which elements are included in a coordinate structure, and what role each plays in the relevant predicate argument structures.

4.3 Verb-particle expressions

Another phenomenon exhibited several times in these ten sentences involves verb-particle expressions, as with *thrash out* and perhaps also *stop by*. Most of the supplied schemes distinguished this dependency, but some simply treated the particle as a modifier of the verb. It would be desirable to explicitly distinguish in a target annotation the contrast between *stopped a session* and *stopped by a session* without having to hunt around in the annotation to see if there happens to be a modifier of *stop* that would dramaticaly change its meaning.

The example with *stop by a session* also highlights the need for an annotation scheme which localizes the differences between competing analyses where possible. Though all of the supplied annotations treat *by* as a particle just like *up* in *"look up the answer"*, in fact *by* fails the clearest test for being a particle, namely the ability to appear after the NP argument: *"*He stopped the session by."* An analysis treating *"by the session"* as a selected-for PP with a semantically empty *by* might better fit the linguistic facts, but the target annotation could remain neutral about this syntactic debate if it simply recorded the predicate as *stop_by*, taking an NP argument just as is usually done for the complement of *rely* in *"rely on us"*.

4.4 Less frequent phenomena

Since each new phenomenon encountered may well require negotiation in order to arrive at a common target annotation, it will be important to include some provisional annotation for relations that have not yet been negotiated. Even these ten example sentences include a few expressions where there was little or no agreement among the schemes about the annotations, such as *"if not more so"* in sentence 30, or *"to be autographed"* in sentence 216. It would be convenient if the target annotation scheme included a noncommittal representation for some parts of a given sentence explicitly noting the lack of clarity about what the structure should be.

4.5 Productive derivational morphology

It was surprising that only one of the annotation schemes (CONLL08) explicitly annotated the nominal gerund *conducting* in sentence 53 as productively related to the verb *conduct*.. While the issue of derivational morphology is of course a slippery slope, the completely productive gerund-forming process in English should be accommodated in any target annotation scheme, as should a small number of other highly productive and morphologically marked derivational regularities, including participial verbs used as prenominal modifiers, and comparative and superlative adjectives. Including this stemming would provide an informative level of detail in the target annotation, and one which can almost always be readily determined from the syntactic context.

5 Next steps

The existing annotation schemes supplied for this task exhibit substantial common ground in the nature and level of detail of information being recorded, making plausible the idea of investing a modest amount of joint effort to negotiate a common target representation which addresses at least some of the issues identified here. The initial com-

mon target annotation scheme should be one which has the following properties:

- Each existing scheme's annotations can be readily mapped to those of the target scheme via an automatic procedure.
- The annotations appear in compact, humanly readable form as sets of tuples recording either predicate-argument dependencies or properties of entities and events, such as number and tense.
- The inventory of recorded distinctions is rich enough to accommodate most of what any one scheme records, though it may not be a superset of all such distinctions. For example, some scheme might record quantifier scope information, yet the target annotation scheme might not, either because it is not of high priority for most participants, or because it would be difficult to produce consistently in a gold standard.

The primary purposes of such a target annotation scheme should be to facilitate the automatic comparison of results across frameworks, and to support evaluation of results against gold standard analyses expressed in this target scheme. It might also be possible to define the scheme such that the target annotations contain enough information to serve as the basis for some application-level tasks such as reasoning, but the primary design criteria should be to enable detailed comparison of analyses.

References

Bresnan, Joan and Ronald M. Kaplan, 1982. Lexical-Functional Grammar. A Formal System for Grammatical Representation. *The Mental Representation of Grammatical Relations*, ed. Joan Bresnan. MIT Press, Cambridge, MA.

Carroll, John, Edward Briscoe and Antonio Sanfilippo. 1998. Parser evaluation: a survey and a new proposal. *Proceedings of the 1st International Conference on Language Resources and Evaluation*.

Clark, Stephen and James R. Curran. 2003. Log-linear models for wide-coverage CCG parsing. *Proceedings of the 2003 conference on Empirical methods in natural language processing*, pp.97–104.

Clark, Stephen and James R. Curran. 2007. Formalism-Independent Parser Evaluation with CCG and DepBank. *Proceedings of the Association for Computational Linguistics 2007*.

Harrison, P., S. Abney, D. Flickinger, C. Gdaniec, R. Grishman, D. Hindle, B. Ingria, M. Marcus, B. Santorini, , and T. Strzalkowski. 1991. Evaluating syntax performance of parser/grammars of English. *Natural Language Processing Systems Evaluation Workshop*, Technical Report RL- TR-91-6, J. G. Neal and S. M. Walter, eds.

Hockenmaier, Julia. 2003. *Data and Models for Statistical Parsing with Combinatory Categorial Grammar*. Ph.D. thesis, University of Edinburgh.

Marcus, Mitchell P., Beatrice Santorini and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English. The Penn Treebank. *Computational Linguistics* 19:313–330.

Pollard, Carl and Ivan A. Sag. 1994. *Head-Driven Phrase Structure Grammar*. The University of Chicago Press and CSLI Publications, Chicago, IL and Stanford, CA.

Preiss, Judita. 2003. Using Grammatical Relations to Compare Parsers. *Proceedings of the European Association for Computational Linguistics 2003*.

Sagae, Kenji, Yusuke Miyao, Takuya Matsuzaki and Jun'ichi Tsujii. 2008. Challenges in Mapping of Syntactic Representations for Framework-Independent Parser Evaluation. *Proceedings of the Workshop on Automated Syntatic Annotations for Interoperable Language Resources at the 1st International Conference on Global Interoperability for Language Resources*, pp.61–68.

Steedman, Mark. 2000. *The syntactic process*. MIT Press, Cambridge, MA.