Introduction to Frontiers in Corpus Annotation II Pie in the Sky

Adam Meyers
New York University
meyers@cs.nyu.edu

1 Introduction

The *Frontiers in Corpus Annotation* workshops are opportunities to discuss the state of the art of corpus annotation in computational linguistics. Corpus annotation has pushed the enitre field in new directions by providing new task definitions and new standards of analysis. At the first *Frontiers in Corpus Annotation* workshop at *HLT-NAACL 2004* we compared assumptions underlying different annotation projects in light of both multilingual applications and the pursuit of merged representations that incorporate the result of various annotation projects.

Beginning September, 2004, several researchers have been collaborating to produce detailed semantic annotation of two difficult sentences. The effort aimed to produce a single unified representation that goes beyond what may currently be feasible to annotate consistently or to generate automatically. Rather this "pie in the sky" annotation effort was an attempt at defining a future goal for semantic analysis. We decided to use the "Pie in the Sky" annotation effort (http://nlp.cs.nyu.edu/meyers/pie-in-the-sky.html) as a theme for this year's workshop. Consequently this theme has been brought out in many of the papers contained in this volume.

The first 4 papers (Pustejovsky et al., 2005; E. W. Hinrichs and S. Kübler and K. Naumann, 2005; Bies et al., 2005; Dinesh et al., 2005) all discuss some aspect of merging annotation. (Pustejovsky et al., 2005) describes issues that arise for merging argument structures for verbs, nouns and discourse connectives, as well as time and anaphora representations. (E. W. Hinrichs and S. Kübler and K. Naumann, 2005) focuses on the merging of syntactic, morphological, semantic and referential annotation. (E. W. Hinrichs and S. Kübler and K. Naumann, 2005) also points out that the "Pie in the Sky"representation lacks syntactic features. This brings to light an important point of discussion: should linguistic analyses be divided out into separate "levels" corresponding to syntax, morphology, discourse, etc. or should/can a single representation represent all such "levels"? As currently conceived, "Pie in the Sky" is intended to be as "language neutral" as possible – this may make adding a real syntactic level difficult. However, arguably, surface relations can be added on as features to Pie in the Sky, even if we delete or ignore those features for some (e.g., language neutral) purposes. Still, other papers present further difficulties for maintaining a single representation that covers multiple modes of analysis. (Bies et al., 2005) discusses possible conflicts between named entity analyses and syntactic structure and (Dinesh et al., 2005) discusses a conflict between discourse structure and syntactic structure. I think it is reasonable to assume that some such conflicts will be resolvable, e.g., I believe that the named entity conflicts point to shortcomings of the original Penn Treebank analysis. However, the discourse structure/syntactic structure conflicts may be harder to solve. In fact, some annotation projects, e.g., the Prague Dependency Treebank (Hajiĉová and Ceplová, 2000), assume that multiple analyses or "levels" are necessary to describe the full range of phenomena.

The 5th through 7th papers (Inui and Okumura, 2005; Calhoun et al., 2005; Wilson and Wiebe, 2005) investigate some additional types of annotation that were not part of the distributed version of Pie in the Sky, but which could be added in principle. In fact, with help from the authors of (Calhoun et al., 2005), I did incorporate their analysis into the latest version (number 6) of the"Pie in the Sky" annotation. Furthermore, it turns out that some units of Information Structure cross the boundaries of the syntactic/semantic constituents, thus raising the sort of difficulties discussed in the previous paragraph. Specifically, information structure divides sentences into themes and rhemes. For the sample two sentences, the rheme boundaries do correspond to syntactic units, but the theme boundaries cross syntactic boundaries, forming units made up of parts of multiple syntactic constituents.

(Palmer et al., 2005; Xue, 2005) (the eighth and

Proceedings of the Workshop on Frontiers in Corpus Annotation II: Pie in the Sky, pages 1–4,
Ann Arbor, June 2005.

eleventh papers) make comparisons of annotated phenomena across English and Chinese. It should be pointed out that seven of the papers at this workshop are predominantly about the annotation of English, one is about German annotation and one is about Japanese annotation. These two are the only papers at the workshop that explicitly discuss attempts to apply the same annotation scheme across two languages.

(McShane et al., 2005; Poesio and Artstein, 2005) (the ninth and tenth papers) both pertain to issues about improving the annotation process. (Poesio and Artstein, 2005) discusses some better ways of assessing inter-annotator agreement, particularly when there is a gray area between correct and incorrect annotation. (McShane et al., 2005) discusses the issue of human-aided annotation (human correction of a machine-generated analysis) as it pertains to a single-integrated annotation scheme, similar in many ways to "Pie in the Sky", although it has been in existence for a lot longer.

2 Issues for Discussion

These papers raise a number of important issues for discussion, some of which I have already touched on.

Question 1: Should the community annotate lots of individual phenomena independently of one another or should we assume an underlying framework and perform all annotation tasks so they are compatible with that framework?

Some of the work presented describes the annotation of fairly narrow linguistic phenomena. Pie in the Sky can be viewed as a framework for unifying these annotation schemata into a single representation (a Unified Linguistic Annotation framework in the sense of (Pustejovsky et al., 2005)). Other work presented assumes that the integrated framework is the object of the annotation rather than the result of merging annotations (E. W. Hinrichs and S. Kübler and K. Naumann, 2005; McShane et al., 2005). There are pros and cons to both approaches.

When researchers decide to annotate one small piece of linguistic analysis (verb argument structure, noun argument structure, coreference, discourse structure, etc.), this has the following potential advantages: (1) exploring one phenomenon in depth may provide a better characterization of that phenomenon. If individual phenomena are examined with this level of care, perhaps we will end up with a better overall analysis; (2) a very focused task definition for the annotator may improve interannotator agreement; and (3) it is sometimes easier to analyze a phenomenon in isolation, especially if there is not a large literature of previous work about it – indeed, trying to integrate this new phenomenon before adequately understanding it may unduly bias one's research. However, by ignoring a more complete theory, these annotation projects run the risk of task-based biases, e.g., classifying predication as coreference or coreference as argument-hood. While an underlying all-inclusive theory could be a useful roadmap, unifying the results of several annotation efforts (and resolving inconsistencies) may yield the same result (as suggested in (Pustejovsky et al., 2005)) while maintaining the advantages of investigating the phenomena separately. On the other hand, as this merging process has not come to completion yet, the jury is still out.

Let's say that, for the sake of argument, the reader accepts the research program where individual annotation efforts are slowly merged into one "Pie in the Sky" type system. There is still another obvious question that arises:

Question 2: Why make up a brand new system like "Pie in the Sky" when there are so many existing frameworks around? For example, Head-Driven Phrase Structure Grammar (Pollard and Sag, 1994) assumes a fairly large feature structure that would seem to accommodate every possible level of linguistic analysis (although in practice most authors in that framework only work on the syntactic and semantic portion of that feature structure).

Our initial motivation for starting fresh is that we wanted the framework to use the minimal features necessary to represent the input annotation systems and to extend them as much as possible. In addition, part of the experiment was an aim to keep features in a somewhat language-neutral form and it is not clear that there are existing frameworks that both share this bias and are sufficiently expressive for our purposes. However, ultimately it might be beneficial to convert "Pie in the Sky" to one or more pre-existing frameworks.

So far, we have limited the scope of "Pie in the Sky" to semantic and (recently) some discourse information as well. However, there are some cases where we found it necessary to include syntactic information, e.g., although heads are semantic arguments of adjective modifiers, the surface relation between the head of the noun phrase and its constituents is important for determining other parts of meaning. For example, although *explosive* would bear the same argument relation to *powerful* in both (a) *The explosive is powerful* and (b) *the powerful explosive*, the interpretation of (b) requires that *powerful* be part of the same unit as *explosive*, e.g., for the proper interpretation of *He bought a powerful explosive*. Thus it may seem like a good idea to ultimately fill out "Pie in the Sky" into a larger framework. However, we would still want to be able to pick out the language-neutral components of the analysis from the language-specific ones.

Question 3: D. Farwell, a member of the workshop committee, has pointed out that there are levels within semantics. The question is how should these multiple levels be handled? The annotated examples did not include phenomena such as metaphor, metonymy or idiomaticity that may have multiple interpretations: literal and intended.

For example, an adequate interpretation of *I love listening to Mozart* would require *Mozart* to be decomposed into *music by Mozart* (although arguably the representation of some of the complex discourse references were of this flavor).

3 What's in the Latest Pie in the Sky Analysis

As of this writing, the latest "Pie in the Sky" analysis includes: (1) argument structure of all parts of speech (verbs, nouns, adjectives, determiners, conjunctions, etc.) using the PropBank/NomBank/Discourse Treebank argument labels (ARG0, ARG1, ARG2, ...), reminiscent of Relational Grammar of the 1970s and 1980s (Perlmutter, 1984), (2) some more specifically labeled FrameNet (Baker et al., 1998) roles for these same constituents; (3) morphological and part of speech features; (4) pointers to gazetteers, both real and hypothetical (thanks to B. Sundheim); (5) Veracity/According-To features based on NYU's proposed FactBank annotation scheme; (6) various coreference features including some based on a proposed extension to NomBank; (7) temporal features based on Timex2 (Ferro et al., 2002) and TimeML (Pustejovsky et al., 2004); and (8) Information Structure features based on (Calhoun et al., 2005). For more detail, please see: http://nlp.cs.nyu.edu/meyers/pie-in-the-sky.html

4 The Future of "Pie in the Sky"

After this workshop, we plan to retire the current two "Pie in the Sky" sentences and start again with some new text. I observed the following obstacles during this experiment: (1) annotation projects were somewhat hesitant to volunteer their time (so we are extremely grateful to all projects that did so.); (2) the target material was not long enough for some annotation approaches to be able to really make their mark, e.g., two sentences are not so interesting for discourse purposes.; and (3) partially due to its length, some interesting phenomena were not well-represented (idioms, metonymy, etc.)

The lack of volunteers may, in part, be related to the scale of the project. We built the project up slowly and invited people to join in, rather than posting a request for annotations to an international list. Initially, this was necessary just to make the project possible to manage. Additionally, inadequacies of the data were probably barriers for projects that focused on discourse phenomena or phenomena that was not well-represented by our data. Nevertheless, using more data may place too heavy a burden on annotation projects and this could make projects hesitant to participate.

With these issues in mind, I note that several sites annotated two longer documents for the recent U.S. Government sponsored Semantic Annotation Planning Meeting at the University of Maryland. This success was, in part, due to the chance for annotation sites to attract government interest in funding their projects. While we will not attempt to duplicate this workshop, I believe that there is an underlying issue that is very important. The field really needs a single test corpus for all new annotation projects.

This test corpus would meet a number of important needs of the annotation community: (1) it would provide a testbed for new annotation schemata; (2) it would provide a large corpus that is annotated in a fairly complete framework – this way focused annotation projects may be able to more easily write specifications in light of where their particular set of phenomena fit into some larger framework; and (3) it would provide a steady flow of input annotation in order to produce a single unified annotation framework.

To make this idea a reality, we need to obtain a consensus on what people would like to annotate. Additionally, we need volunteers to translate this same corpus into other languages, as we would inevitably choose an English corpus. Of course, if we could find a suitable text that was already translated in multiple languages, this would save time. The perfect text would be article length (loosely defined); include difficult to handle phenomena (idioms, metonymy, etc.); include a wide range of annotatable linguistic phenomena and not have copyright restrictions which would hamper the project. It would, of course, be helpful if the annotation community would provide input on which text to choose – this would avoid a situation where one could not annotate the test text because the target phenomenon is not represented there.

In summary, I have used this introduction to both summarize how the papers of this workshop fit together, to propose some unifying themes for discussion, and to propose an agenda for how to proceed after the workshop is over. We hope to see some of these ideas come to fruition before "Frontiers in Corpus Annotation III."

References

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet Project. In *Proceedings of Coling-ACL98: The 17th International Conference on Computational Linguistics and the 36th Meeting of the Association for Computational Linguistics*, pages 86–90.

A. Bies, S. Kulick, and M. Mandel. 2005. Parallel Entity and Treebank Annotation. In *ACL 2005 Workshop: Frontiers in Corpus Annotation II: Pie in the Sky*.

S. Calhoun, M. Nissim, M. Steedman, and J. Brenier. 2005. A Framework for Annotating Information Struc-

ture in Discourse. In *ACL 2005 Workshop: Frontiers in Corpus Annotation II: Pie in the Sky*.

N. Dinesh, A. Lee, E. Miltsakaki, R. Prasad, A. Joshi, and B. Webber. 2005. Attribution and the (Non-)Alignment of Syntactic and Discourse Arguments of Connectives. In *ACL 2005 Workshop: Frontiers in Corpus Annotation II: Pie in the Sky*.

E. W. Hinrichs and S. Kübler and K. Naumann. 2005. A Unified Reprewsentation for Morphological, Syntactic, Semantic, and Referential Annotations. In *ACL 2005 Workshop: Frontiers in Corpus Annotation II: Pie in the Sky*.

L. Ferro, L. Gerber, I. Mani, B. Sundheim, and G. Wilson. 2002. Instruction Manual for the Annotation of Temporal Expressions. MITRE Washington C3 Center, McLean, Virginia.

Eva Hajiĉová and Mark'eta Ceplová. 2000. Deletions and Their Reconstruction in Tectogrammatical Syntactic Tagging of Very Large Corpora. In *Proceedings of Coling 2000: The 18th International Conference on Computational Linguistics*, pages 278–284.

T. Inui and M. Okumura. 2005. Investigating the Characteristics of Causal Relations in Japanese Text. In *ACL 2005 Workshop: Frontiers in Corpus Annotation II: Pie in the Sky*.

M. McShane, S. Nirenburg, S. Beale, and T. O'Hara. 2005. Semantically Rich Human-Aided Machine Annotation. In *ACL 2005 Workshop: Frontiers in Corpus Annotation II: Pie in the Sky*.

M. Palmer, N. Xue, O. Babko-Malaya, J. Chen, and B. Snyder. 2005. A Parallel Proposition Bank II for Chinese and English. In *ACL 2005 Workshop: Frontiers in Corpus Annotation II: Pie in the Sky*.

David. M. Perlmutter. 1984. *Studies in Relational Grammar 1*. The University of Chicago Press, Chicago.

M. Poesio and R. Artstein. 2005. The Reliability of Anaphoric Annotation, Reconsidered: Taking Ambiguity into Account. In *ACL 2005 Workshop: Frontiers in Corpus Annotation II: Pie in the Sky*.

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

J. Pustejovsky, B. Ingria, R. Sauri, J. Castano, J. Littman, R. Gaizauskas, A. Setzer, G. Katz, and I. Mani. 2004. The Specification Language TimeML. In I. Mani, J. Pustejovsky, and R. Gaizauskas, editors, *The Language of Time: A Reader*. Oxford University Press, Oxford.

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*.

T. Wilson and J. Wiebe. 2005. Annotating Attributions and Private States. In *ACL 2005 Workshop: Frontiers in Corpus Annotation II: Pie in the Sky*.

N. Xue. 2005. Annotating Discourse Connectives in the Chinese Treebank. In *ACL 2005 Workshop: Frontiers in Corpus Annotation II: Pie in the Sky*.