The Manually Annotated Sub-Corpus: A Community Resource For and By the People

Nancy Ide
Department of Computer Science
Vassar College
Poughkeepsie, NY, USA
ide@cs.vassar.edu

Collin Baker
International Computer Science Institute
Berkeley, California USA
collinb@icsi.berkeley.edu

Christiane Fellbaum
Princeton University
Princeton, New Jersey USA
fellbaum@princeton.edu

Rebecca Passonneau
Columbia University
New York, New York USA
becky@cs.columbia.edu

Abstract

The Manually Annotated Sub-Corpus (MASC) project provides data and annotations to serve as the base for a community-wide annotation effort of a subset of the American National Corpus. The MASC infrastructure enables the incorporation of contributed annotations into a single, usable format that can then be analyzed as it is or ported to any of a variety of other formats. MASC includes data from a much wider variety of genres than existing multiply-annotated corpora of English, and the project is committed to a fully open model of distribution, without restriction, for all data and annotations produced or contributed. As such, MASC is the first large-scale, open, community-based effort to create much needed language resources for NLP. This paper describes the MASC project, its corpus and annotations, and serves as a call for contributions of data and annotations from the language processing community.

1 Introduction

The need for corpora annotated for multiple phenomena across a variety of linguistic layers is keenly recognized in the computational linguistics community. Several multiply-annotated corpora exist, especially for Western European languages and for spoken data, but, interestingly, broad-based English language corpora with robust annotation for diverse linguistic phenomena are relatively rare. The most widely-used corpus of English, the British National Corpus, contains only part-of-speech annotation; and although it contains a wider range of annotation types, the fifteen million word Open American National Corpus annotations are largely unvalidated. The most well-known multiply-annotated and validated corpus of English is the one million word *Wall Street Journal* corpus known as the Penn Treebank (Marcus et al., 1993), which over the years has been fully or partially annotated for several phenomena over and above the original part-of-speech tagging and phrase structure annotation. The usability of these annotations is limited, however, by the fact that many of them were produced by independent projects using their own tools and formats, making it difficult to combine them in order to study their inter-relations. More recently, the OntoNotes project (Pradhan et al., 2007) released a one million word English corpus of newswire, broadcast news, and broadcast conversation that is annotated for Penn Treebank syntax, PropBank predicate argument structures, coreference, and named entities. OntoNotes comes closest to providing a corpus with multiple layers of annotation that can be analyzed as a unit via its representation of the annotations in a "normal form". However, like the *Wall Street Journal* corpus, OntoNotes is limited in the range of genres it includes. It is also limited to only those annotations that may be produced by members of the OntoNotes project. In addition, use of the data and annotations with software other than the OntoNotes database API is not necessarily straightforward.

The sparseness of reliable multiply-annotated corpora can be attributed to several factors. The greatest obstacle is the high cost of manual production and validation of linguistic annotations. Furthermore, the production and annotation of corpora, even when they involve significant scientific research, often do not, *per se*, lead to publishable research results. It is therefore understand-

able that many research groups are unwilling to get involved in such a massive undertaking for relatively little reward.

The Manually Annotated Sub-Corpus (MASC) (Ide et al., 2008) project has been established to address many of these obstacles to the creation of large-scale, robust, multiply-annotated corpora. The project is providing appropriate data and annotations to serve as the base for a community-wide annotation effort, together with an infrastructure that enables the representation of internally-produced and contributed annotations in a single, usable format that can then be analyzed as it is or ported to any of a variety of other formats, thus enabling its immediate use with many common annotation platforms as well as off-the-shelf concordance and analysis software. The MASC project's aim is to offset some of the high costs of producing high quality linguistic annotations via a distribution of effort, and to solve some of the usability problems for annotations produced at different sites by harmonizing their representation formats.

The MASC project provides a resource that is significantly different from OntoNotes and similar corpora. It provides data from a much wider variety of genres than existing multiply-annotated corpora of English, and all of the data in the corpus are drawn from current American English so as to be most useful for NLP applications. Perhaps most importantly, the MASC project is committed to a fully open model of distribution, without restriction, for all data and annotations. It is also committed to incorporating diverse annotations contributed by the community, regardless of format, into the corpus. As such, MASC is the first large-scale, open, community-based effort to create a much-needed language resource for NLP. This paper describes the MASC project, its corpus and annotations, and serves as a call for contributions of data and annotations from the language processing community.

2 MASC: The Corpus

MASC is a balanced subset of 500K words of written texts and transcribed speech drawn primarily from the Open American National Corpus (OANC)[1]. The OANC is a 15 million word (and growing) corpus of American English produced since 1990, all of which is in the public domain or otherwise free of usage and redistribution restrictions.

Genre	No. texts	Total words
Email	2	468
Essay	4	17516
Fiction	4	20413
Gov't documents	1	6064
Journal	10	25635
Letters	31	10518
Newspaper/newswire	41	17951
Non-fiction	4	17118
Spoken	11	25783
Debate transcript	2	32325
Court transcript	1	20817
Technical	3	15417
Travel guides	4	12463
Total	118	222488

Table 1: MASC Composition (first 220K)

Where licensing permits, data for inclusion in MASC is drawn from sources that have already been heavily annotated by others. So far, the first 80K increment of MASC data includes a 40K subset consisting of OANC data that has been previously annotated for PropBank predicate argument structures, Pittsburgh Opinion annotation (opinions, evaluations, sentiments, etc.), TimeML time and events[2], and several other linguistic phenomena. It also includes a handful of small texts from the so-called Language Understanding (LU) Corpus[3] that has been annotated by multiple groups for a wide variety of phenomena, including events and committed belief. All of the first 80K increment is annotated for Penn Treebank syntax. The second 120K increment includes 5.5K words of *Wall Street Journal* texts that have been annotated by several projects, including Penn Treebank, PropBank, Penn Discourse Treebank, TimeML, and the Pittsburgh Opinion project. The composition of the 220K portion of the corpus annotated so far is shown in Table 1. The remaining 280K of the corpus fills out the genres that are under-represented in the first portion and includes a few additional genres such as blogs and tweets.

3 MASC Annotations

Annotations for a variety of linguistic phenomena, either manually produced or corrected from output of automatic annotation systems, are being added

[1] http://www.anc.org

[2] The TimeML annotations of the data are not yet completed.

[3] MASC contains about 2K words of the 10K LU corpus, eliminating non-English and translated LU texts as well as texts that are not free of usage and redistribution restrictions.

Annotation type	Method	No. texts	No. words
Token	Validated	118	222472
Sentence	Validated	118	222472
POS/lemma	Validated	118	222472
Noun chunks	Validated	118	222472
Verb chunks	Validated	118	222472
Named entities	Validated	118	222472
FrameNet frames	Manual	21	17829
HSPG	Validated	40*	30106
Discourse	Manual	40*	30106
Penn Treebank	Validated	97	87383
PropBank	Validated	92	50165
Opinion	Manual	97	47583
TimeBank	Validated	34	5434
Committed belief	Manual	13	4614
Event	Manual	13	4614
Coreference	Manual	2	1877

Table 2: Current MASC Annotations (* projected)

to MASC data in increments of roughly 100K words. To date, validated or manually produced annotations for 222K words have been made available.

The MASC project is itself producing annotations for portions of the corpus for WordNet senses and FrameNet frames and frame elements. To derive maximal benefit from the semantic information provided by these resources, the entire corpus is also annotated and manually validated for shallow parses (noun and verb chunks) and named entities (person, location, organization, date and time). Several additional types of annotation have either been contracted by the MASC project or contributed from other sources. The 220K words of MASC I and II include seventeen different types of linguistic annotation[4], shown in Table 2.

All MASC annotations, whether contributed or produced in-house, are transduced to the Graph Annotation Framework (GrAF) (Ide and Suderman, 2007) defined by ISO TC37 SC4's Linguistic Annotation Framework (LAF) (Ide and Romary, 2004). GrAF is an XML serialization of the LAF abstract model of annotations, which consists of a directed graph decorated with feature structures providing the annotation content. GrAF's primary role is to serve as a "pivot" format for transducing among annotations represented in different formats. However, because the underlying data structure is a graph, the GrAF representation itself can serve as the basis for analysis via application of graph-analytic algorithms such as common subtree detection.

The layering of annotations over MASC texts dictates the use of a stand-off annotation representation format, in which each annotation is contained in a separate document linked to the primary data. Each text in the corpus is provided in UTF-8 character encoding in a separate file, which includes no annotation or markup of any kind. Each file is associated with a set of GrAF standoff files, one for each annotation type, containing the annotations for that text. In addition to the annotation types listed in Table 2, a document containing annotation for logical structure (titles, headings, sections, etc. down to the level of paragraph) is included. Each text is also associated with (1) a header document that provides appropriate metadata together with machine-processable information about associated annotations and interrelations among the annotation layers; and (2) a segmentation of the primary data into minimal regions, which enables the definition of different tokenizations over the text. Contributed annotations are also included in their original format, where available.

3.1 WordNet Sense Annotations

A focus of the MASC project is to provide corpus evidence to support an effort to harmonize sense distinctions in WordNet and FrameNet (Baker and Fellbaum, 2009), (Fellbaum and Baker, to appear). The WordNet and FrameNet teams have selected for this purpose 100 common polysemous words whose senses they will study in detail, and the MASC team is annotating occurrences of these words in the MASC. As a first step, fifty occurrences of each word are annotated using the WordNet 3.0 inventory and analyzed for problems in sense assignment, after which the WordNet team may make modifications to the inventory if needed. The revised inventory (which will be released as part of WordNet 3.1) is then used to annotate 1000 occurrences. Because of its small size, MASC typically contains less than 1000 occurrences of a given word; the remaining occurrences are therefore drawn from the 15 million words of the OANC. Furthermore, the FrameNet team is also annotating one hundred of the 1000 sentences for each word with FrameNet frames and frame elements, providing direct comparisons of WordNet and FrameNet sense assignments in

[4]This includes WordNet sense annotations, which are not listed in Table 2 because they are not applied to full texts; see Section 3.1 for a description of the WordNet sense annotations in MASC.

attested sentences.[5]

For convenience, the annotated sentences are provided as a stand-alone corpus, with the WordNet and FrameNet annotations represented in standoff files. Each sentence in this corpus is linked to its occurrence in the original text, so that the context and other annotations associated with the sentence may be retrieved.

3.2 Validation

Automatically-produced annotations for sentence, token, part of speech, shallow parses (noun and verb chunks), and named entities (person, location, organization, date and time) are hand-validated by a team of students. Each annotation set is first corrected by one student, after which it is checked (and corrected where necessary) by a second student, and finally checked by both automatic extraction of the annotated data and a third pass over the annotations by a graduate student or senior researcher. We have performed inter-annotator agreement studies for shallow parses in order to establish the number of passes required to achieve near-100% accuracy.

Annotations produced by other projects and the FrameNet and Penn Treebank annotations produced specifically for MASC are semi-automatically and/or manually produced by those projects and subjected to their internal quality controls. No additional validation is performed by the ANC project.

The WordNet sense annotations are being used as a base for an extensive inter-annotator agreement study, which is described in detail in (Passonneau et al., 2009), (Passonneau et al., 2010). All inter-annotator agreement data and statistics are published along with the sense tags. The release also includes documentation on the words annotated in each round, the sense labels for each word, the sentences for each word, and the annotator or annotators for each sense assignment to each word in context. For the multiply annotated data in rounds 2-4, we include raw tables for each word in the form expected by Ron Artstein's calculate_alpha.pl perl script[6], so that the agreement numbers can be regenerated.

[5]Note that several MASC texts have been fully annotated for FrameNet frames and frame elements, in addition to the WordNet-tagged sentences.

[6]http://ron.artstein.org/resources/calculate-alpha.perl

4 MASC Availability and Distribution

Like the OANC, MASC is distributed without license or other restrictions from the American National Corpus website[7]. It is also available from the Linguistic Data Consortium (LDC)[8] for a nominal processing fee.

In addition to enabling download of the entire MASC, we provide a web application that allows users to select some or all parts of the corpus and choose among the available annotations via a web interface (Ide et al., 2010). Once generated, the corpus and annotation bundle is made available to the user for download. Thus, the MASC user need never deal directly with or see the underlying representation of the stand-off annotations, but gains all the advantages that representation offers. The following output formats are currently available:

1. in-line XML (XCES[9]), suitable for use with the BNCs XAIRA search and access interface and other XML-aware software;
2. token / part of speech, a common input format for general-purpose concordance software such as MonoConc[10], as well as the Natural Language Toolkit (NLTK) (Bird et al., 2009);
3. CONLL IOB format, used in the Conference on Natural Language Learning shared tasks.[11]

5 Tools

The ANC project provides an API for GrAF annotations that can be used to access and manipulate GrAF annotations directly from Java programs and render GrAF annotations in a format suitable for input to the open source GraphViz[12] graph visualization application.[13] Beyond this, the ANC project does not provide specific tools for use of the corpus, but rather provides the data in formats suitable for use with a variety of available applications, as described in section 4, together with means to import GrAF annotations into major annotation software platforms. In particular, the ANC project provides plugins for the General

[7]http://www.anc.org

[8]http://www.ldc.upenn.edu

[9]XML Corpus Encoding Standard, http://www.xces.org

[10]http://www.athel.com/mono.html

[11]http://ifarm.nl/signll/conll

[12]http://www.graphviz.org/

[13]http://www.anc.org/graf-api

Architecture for Text Engineering (GATE) (Cunningham et al., 2002) to input and/or output annotations in GrAF format; a "CAS Consumer" to enable using GrAF annotations in the Unstructured Information Management Architecture (UIMA) (Ferrucci and Lally, 2004); and a corpus reader for importing MASC data and annotations into NLTK[14].

Because the GrAF format is isomorphic to input to many graph-analytic tools, existing graph-analytic software can also be exploited to search and manipulate MASC annotations. Trivial merging of GrAF-based annotations involves simply combining the graphs for each annotation, after which graph minimization algorithms[15] can be applied to collapse nodes with edges to common subgraphs to identify commonly annotated components. Graph-traversal and graph-coloring algorithms can also be applied in order to identify and generate statistics that could reveal interactions among linguistic phenomena that may have previously been difficult to observe. Other graph-analytic algorithms — including common sub-graph analysis, shortest paths, minimum spanning trees, connectedness, identification of articulation vertices, topological sort, graph partitioning, etc. — may also prove to be useful for mining information from a graph of annotations at multiple linguistic levels.

6 Community Contributions

The ANC project solicits contributions of annotations of any kind, applied to any part or all of the MASC data. Annotations may be contributed in any format, either inline or standoff. All contributed annotations are ported to GrAF standoff format so that they may be used with other MASC annotations and rendered in the various formats the ANC tools generate. To accomplish this, the ANC project has developed a suite of internal tools and methods for automatically transducing other annotation formats to GrAF and for rapid adaptation of previously unseen formats.

Contributions may be emailed to `anc@cs.vassar.edu` or uploaded via the ANC website[16]. The validity of annotations and supplemental documentation (if appropriate) are the responsibility of the contributor. MASC users may contribute evaluations and error reports for the various annotations on the ANC/MASC wiki[17].

Contributions of unvalidated annotations for MASC and OANC data are also welcomed and are distributed separately. Contributions of unencumbered texts in any genre, including stories, papers, student essays, poetry, blogs, and email, are also solicited via the ANC web site and the ANC FaceBook page[18], and may be uploaded at the contribution page cited above.

7 Conclusion

MASC is already the most richly annotated corpus of English available for widespread use. Because the MASC is an open resource that the community can continually enhance with additional annotations and modifications, the project serves as a model for community-wide resource development in the future. Past experience with corpora such as the *Wall Street Journal* shows that the community is eager to annotate available language data, and we anticipate even greater interest in MASC, which includes language data covering a range of genres that no existing resource provides. Therefore, we expect that as MASC evolves, more and more annotations will be contributed, thus creating a massive, inter-linked linguistic infrastructure for the study and processing of current American English in its many genres and varieties. In addition, by virtue of its WordNet and FrameNet annotations, MASC will be linked to parallel WordNets and FrameNets in languages other than English, thus creating a global resource for multi-lingual technologies, including machine translation.

Acknowledgments

The MASC project is supported by National Science Foundation grant CRI-0708952. The WordNet-FrameNet alignment work is supported by NSF grant IIS 0705155.

References

Collin F. Baker and Christiane Fellbaum. 2009. WordNet and FrameNet as complementary resources for annotation. In *Proceedings of the Third Linguistic*

[14] Available in September, 2010.

[15] Efficient algorithms for graph merging exist; see, e.g., (Habib et al., 2000).

[16] http://www.anc.org/contributions.html

[17] http://www.anc.org/masc-wiki

[18] http://www.facebook.com/pages/American-National-Corpus/42474226671

Annotation Workshop, pages 125–129, Suntec, Singapore, August. Association for Computational Linguistics.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*. O'Reilly Media, 1st edition.

Hamish Cunningham, Diana Maynard, Kalina Bontcheva, and Valentin Tablan. 2002. GATE: A framework and graphical development environment for robust nlp tools and applications. In *Proceedings of ACL'02*.

Christiane Fellbaum and Collin Baker. to appear. Aligning verbs in WordNet and FrameNet. *Linguistics*.

David Ferrucci and Adam Lally. 2004. UIMA: An architectural approach to unstructured information processing in the corporate research environment. *Natural Language Engineering*, 10(3-4):327–348.

Michel Habib, Christophe Paul, and Laurent Viennot. 2000. Partition refinement techniques: an interesting algorithmic tool kit. *International Journal of Foundations of Computer Science*, 175.

Nancy Ide and Laurent Romary. 2004. International standard for a linguistic annotation framework. *Natural Language Engineering*, 10(3-4):211–225.

Nancy Ide and Keith Suderman. 2007. GrAF: A graph-based format for linguistic annotations. In *Proceedings of the Linguistic Annotation Workshop*, pages 1–8, Prague, Czech Republic, June. Association for Computational Linguistics.

Nancy Ide, Collin Baker, Christiane Fellbaum, Charles Fillmore, and Rebecca Passonneau. 2008. MASC: The Manually Annotated Sub-Corpus of American English. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC)*, Marrakech, Morocco.

Nancy Ide, Keith Suderman, and Brian Simms. 2010. ANC2Go: A web application for customized corpus creation. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC)*, Valletta, Malta, May. European Language Resources Association.

Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Rebecca J. Passonneau, Ansaf Salleb-Aouissi, and Nancy Ide. 2009. Making sense of word sense variation. In *SEW '09: Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions*, pages 2–9, Morristown, NJ, USA. Association for Computational Linguistics.

Rebecca Passonneau, Ansaf Salleb-Aouissi, Vikas Bhardwaj, and Nancy Ide. 2010. Word sense annotation of polysemous words by multiple annotators. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC)*, Valletta, Malta.

Sameer S. Pradhan, Eduard Hovy, Mitch Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2007. OntoNotes: A unified relational semantic representation. In *ICSC '07: Proceedings of the International Conference on Semantic Computing*, pages 517–526, Washington, DC, USA. IEEE Computer Society.