From Linguistic Annotations to Knowledge Objects

Bonnie J. Dorr (University of Maryland and HLT Center of Excellence, bonnie@umiacs.umd.edu)
Saif Mohammad (University of Maryland, saif@umiacs.umd.edu)
Boyan A. Onyshkevych (Department of Defense, boyan.o@gmail.com)

An important step in the development of systems that require deeper knowledge (e.g., systems that might ultimately infer person-person relationships, event-event relationships, sentiment, and other important information about the world) is the mapping between linguistic annotations that have been developed by researchers and the knowledge objects that might be derived from those linguistic annotations. The difference between linguistic annotations and knowledge objects is that the former are tags on raw data, whereas the latter are representational entities that are derived beyond the actual surface tokens in the raw data.

At HLT Center of Excellence, we are in the defining atomic units (both at the linguistic level and at the deeper knowledge level) that have not been captured entirely in earlier annotation efforts. In our preliminary investigation the linguistic annotations for which we are defining APIs are as follows (based on a wide range of existing annotation schemes in the literature including, but not limited to [1,2,3,4,5,6,7,8,9,10,11]):

> *Arg, Aspect, Assertion, Attribute Structure, Communicative Opposition, Committed Belief, Comparison, Complex Coref, Event Coref, Simple Coref, Dialog act, Dialog Function Unit, Dialog Link, Discourse Relation, Entity, Event Status, Lexical Relation, Modality, Morphological Class, Predicate, Quantity, Referential Type, Rhetorical Device, Sentence Structure, Social Register, Temporal Connective, Thematic Role, Time, Word Sense.*

The knowledge objects for which we are defining APIs are as follows (based on a wide range of existing annotation schemes in the literature including, but not limited to [1,2,5,12,13,14]):

> *Composite Event, Condition, Confidence, Deontic, Epistemic, Epiteuctic, Evaluative, Event Object, Event-Event Relation, Intent, Location Object, Organization Object, Other Attribute, Other Entity Relation, Permission, Person Object, Personal Attribute, Person-Person Relationship, Philosophy, Potential, Quality, Role, State of Affairs, Time object, Timeline.*

An example of the instantiation of these two types of APIs for a given sentence is provided in the Appendix. Once the task of defining (and ultimately implementing) these APIs is complete, it is our intention to develop prototype "language understanding" modules that take the output of an automatic tagger of linguistic information and produce these knowledge objects.

We expect that the automatically produced knowledge objects will be crucial for language analysis systems and will improve the performance of those systems. In addition, confidence values—--an expected output of such systems-—-—are a critical aspect of information that may enable a more focused analysis of incoming data.

REFERENCES (Note: References 1-8 below can be found at the following website: https://intranet.umiacs.umd.edu/conferences/sapm05/Annotation-Utility.html)

[1] Davis, Levin, Mitamura, Annotating Event Coreference, March 9, 2008
[2] Diab, Rambow, Annotating Committed Belief, March 25, 2008
[3] Habash, Ramshaw, Temporal Parsing Manual, January, 2008
[4] Mohammad, Dorr, Hirst, EMNLP-2008
[5] Nirenburg, McShane, Annotating Modality, March 19, 2008
[6] Passonneau, Rambow: Annotating Dialog Structure, March 25, 2008
[7] Ramshaw, Proposed Annotation Directions, May, 2008;
[8] Ramshaw, Coreference Involving Complex Entities, January 31, 2008
[9] Mitchell, Strassel, ACE Corpus Development, 2002
(http://projects.ldc.upenn.edu/Projects/LDC_Institute/Mitchell/**ACE**_LDC_06272002.ppt)
[10] Palmer, Proposition Bank: A Resource of Predicate-Argument Relations, 2001.
(http://www.authorstream.com/Presentation/aksu-28806-palmer-Proposition-Bank-resource-predicate-argument-relations-Outline-Generalizing-Sentences-as-Entertainment-ppt-powerpoint/)
[11] Pradhan, Hovy, Marcus, Palmer, Ramshaw, Weischedel: OntoNotes: A Unified Relational Semantic Representation. ICSC 2007. (http://www.isi.edu/natural-language/people/hovy/papers/07ICSC-OntoNotes-database.pdf)

[12] Prasad, Miltsakaki, Dinesh, Lee, Joshi, Robaldo, Webber, Penn Discourse Treebank 2.0 Annotation Manual, 2007. (http://www.seas.upenn.edu/~pdtb/PDTBAPI/pdtb-annotation-manual.pdf)
[13] Timex2 QuickGuide, November 2003 (http://fofoca.mitre.org/annotation_guidelines/timex2-quick-guide-2003.pdf)
[14] Weibe, Janyce, Corpus and Opinion Recognition System, 2005. (http://www.cs.pitt.edu/mpqa/)

ACKNOWLEDGMENTS

This work is supported, in part, by the National Science Foundation under Grants No. IIS-0705832 and IIS-0531176, and in part, by the Human Language Technology Center of Excellence. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the sponsor.

APPENDIX

This appendix presents representative linguistic annotations and knowledge units for the sentence *John demanded that Bill put in the order for the purchase to be undertaken by the end of the month*.

LINGUISTIC ANNOTATIONS:
[PRED1: <predType=DEMAND > <startSpan=5> <endSpan=12> <argList=ARG1, ARG2, ARG3>]
[ARG1: <argType=ARG0> <startSpan=0> <endSpan=3>]
[ENT1: <normalizedName=John> <entityType= PER> <startSpan=0> <endSpan=3>]
[THEM ROLE1: <Pred=PRED1> <predOrEntity=ARG1> <roleType= Agent > <startSpan=0> <endSpan=3>]
[ARG2: <argType=ARG1> <startSpan=19> <endSpan=22>]
[ENT2: <normalizedName=Bill> <entityType= PER> <startSpan=19> <endSpan=22>]
[THEM ROLE2: <Pred=PRED1> <predOrEntity=ARG2> <roleType= Theme > <startSpan=19> <endSpan=22>]
[ARG3: <argType=ARG2> <startSpan=24> <endSpan=49>]
[THEM ROLE3: <Pred=PRED1> <predOrEntity=ARG3> <roleType= Proposition> <startSpan=24> <endSpan=39>]
[ARG4: <argType=ARGM> <startSpan=51> <endSpan=END>]
[THEM ROLE4: <Pred=PRED1> <predOrEntity=ARG4> <roleType= Aim> <startSpan=51> <endSpan=END>]
[DIALOG ACT1: <daType = RequestAction> <startSpan=5> <endSpan=35>]
[MODALITY1: <modalityType = Obligative> <startSpan=5> <endSpan=12>]
[COMMITTED BELIEF1: <predOrEntity=PRED1> <status=CB>]
[TIME1: <normalizedTime=[]> <temporalType = S> <inherentTime = S, P> <assertionType=null><startSpan=5> <endSpan=12>]
[TIME2: <normalizedTime=[]> <temporalType = E > <inherentTime = U, F> <assertionType=null><startSpan=24> <endSpan=26>]
[TIME3: <normalizedTime=[]> <temporalType = E> <inherentTime = U, F> <assertionType=null><startSpan=64> <endSpan=73>]
[TIME4: <normalizedTime=[by March 31 2004]> <temporalType = D> <inherentTime = S, F> <assertionType=null> <startSpan=93> <endSpan=97>]
[EVENT STATUS: <Pred1> <status = happened >]

KNOWLEDGE UNITS:
[PERS OBJ1: <PersNameObj1> <AddrObj1> <PersAttrList1>]
[PERS OBJ2: <PersNameObj2> <AddrObj2> <PersAttrList2>]
[EVENT OBJ1: <eventType=kb-ptr-DEMAND> <TimeObj1> <LocObj1> <persObjList1>]
[EVENT OBJ2: <eventType=kb-ptr-PURCHASE> <TimeObj2> <LocObj2> <persObjList2>]
[PERS PERS REL1: <PersObj1> <PersObj2> <relnType = subordinate >]
[DEONTIC1: <value=.9> <scope=EVENT OBJ1> <attributedTo=PERS OBJ1>]
[PERS ATTR1: <ADDR = addrObj1> <POB = LOC OBJ1> <DOB = TIME OBJ1> <ID = 101> <gender=M> <height=183> ...]
[PERS ATTR2: <ADDR = addrObj2> <POB = LOC OBJ2> <DOB = TIME OBJ2> <ID = 102> <gender=M> <height=179> ...]