OntoNotes: The 90% Solution

Sameer S. Pradhan and Nianwen Xue

OntoNotes is a five year multi-site collaboration between BBN Technologies, Information Sciences Institute of University of Southern California, University of Colorado, University of Pennsylvania and Brandeis University. The goal of the OntoNotes project is to provide linguistic data annotated with a skeletal representation of the literal meaning of sentences including syntactic parse, predicate-argument structure, coreference, and word senses linked to an ontology, allowing a new generation of language understanding technologies to be developed with new functional capabilities.

In its third year of existence, the OntoNotes project has generated a large amount of high quality data covering various layers of linguistic annotation. This is probably the first time that data of such quality has been available in large quantities covering multiple genres (newswire, broadcast news, broadcast conversation and weblogs) and languages (English, Chinese and Arabic). The guiding principle has been to find a "sweet spot" in the space of *inter-tagger agreement*, *productivity*, and *depth of representation*. The most effective use of this resource for research requires simultaneous access to multiple layers of annotation. This has been made possible by representing the corpus with a relational database to accommodate the dense connectedness of the data and ensure consistency across layers. In order to facilitate ease of understanding and manipulability, the database has also been supplemented with a object-oriented Python API.

The tutorial consists of two parts. In the first part we will familiarize the user with this new resource, describe the various layers of annotations in some detail and discuss the linguistic principles and sometimes practical considerations behind the important design decisions that shapes the corpus. We will also describe the salient differences between the three languages at each layer of annotation and how linguistic peculiarities of different languages were handled in the data.

In the second part, we will describe the data formats of each of the layers and talk about various design decisions that went into the creation of the architecture of the database and the individual tables comprising it, along with issues that came up during the representation process and compromises that were made without sacrificing some primary objectives one of which being the independent existence of each layer that is necessary to allow multi-site collaboration. We will explain how the database schema attempts to interconnect all the layers. Then we will go into the details of the Python API that allows easy access to each of the layers and show that by making the objects closely resemble database tables, the API allows for their flexible integration. This will be followed by a hands-on working session.

1 Tutorial Outline

1. Annotation Layers
    - Overview of OntoNotes
    - Design principles
        - Depth of annotation
        - Consistency (ITA)
        - Linguistics principles
    - Potential applications
        - Question Answering
        - Machine Translation
    - Layers of Annotation in English, Chinese and Arabic
        - Treebank
        - PropBank
        - Word Sense

Proceedings of NAACL HLT 2009: Tutorials, pages 11–12,
Boulder, Colorado, June 2009. ©2009 Association for Computational Linguistics

        - Name
        - Coreference
        - Ontology
    - Comparison with existing multi-layer annotation corpora

2. Data Access API
    - Data
        - File format
        - Metadata specification
    - Database schema representing each layer of annotation
        - ER diagram
        - Inter-connection between the annotation layers (database tables)
    - Python Access API
        - Introduction to the Python modules
        - Correspondence between MySQL tables and Python classes
        - Introduction to some frequently used module functionalities
        - Extending the API to add a new layer of annotation
    - Hands on Session
        - Creating a sample OntoNotes database from MySQL dump file
        - Loading it into memory
        - Creating Python objects representing various annotation layers
        - Performing cross-layer queries using a combination of API and database
            - We will provide some sample queries
            - Users can use their own experience to generate novel queries
        - Manipulating the data as in Python world and MySQL world
        - Writing the modified versions back to the database

2 Target Audience

This tutorial is designed for people interested in using one or more layers of OntoNotes in their research to further language understanding through improved shallow semantic analysis. Detailed knowledge of any of the layers is not necessary. Some familiarity with Python would be preferable.

Sameer Pradhan is a Research Scientist at BBN Technologies. His research interests include computational semantics, question answering, application of machine learning to language understanding and annotation science. He have been leading the data integration and coreference annotation effort in the DARPA funded GALE OntoNotes project at BBN. In the past he was the technical lead on the AQUAINT project at BBN. He serves on the ACL SIGANN committee and has been one of the organizers of the Linguistics Annotation Workshops (LAW II and III) He has been on the programme committees of Workshop on UIMA for NLP, and Conference on Global Interoperability of Language Resources (ICGL) He has also served on the guest Editorial Board of Computational Linguistics: Special issue on Semantic Role Labeling. He got his PhD in Computer Science at the University of Colorado at Boulder.

Nianwen Xue is an Assistant Professor of Language & Linguistics and Computer Science at Brandeis University. His research interests include formal representation of linguistic structures and its impact on natural language processing, aspects of syntax, computational linguistics, corpus linguistics and Chinese language processing. He is currently leading the effort to expand the Chinese Treebank, Proposition Bank and word sense annotation, funded by DARPA as part of the GALE OntoNotes project. He serves on the ACL SIGANN committee. He is one of the organizers of the Linguistics Annotation Workshops (LAW II and III) and is also on the organizing committee of the CoNLL Shared Task on Syntactic and Semantic Dependencies in Multiple Languages. He has also served on the guest Editorial Board of Computational Linguistics: Special issue on Semantic Role Labeling. He got his PhD in linguistics from University of Delaware.