OntoNotes Release 2.0

with OntoNotes DB Tool v. 0.92 beta

and OntoViewer v. 0.9 beta

http://www.bbn.com/NLP/OntoNotes

2007-11-15

Ralph Weischedel, Sameer Pradhan, Lance Ramshaw, Tricia Khleif

University of Colorado at Boulder

Martha Palmer, Nianwen Xue

Mitchell Marcus, Ann Taylor, Craig Greenberg

Eduard Hovy, Robert Belvin, Ann Houston (from Grammarsmith)

Contents

1 Introduction

This document describes release 2.0 of OntoNotes, an annotated corpus whose development is being supported under the GALE program of the Defense Advanced Research Projects Agency, Contract No. HR0011-06-C-0022. More detailed documents (referred to at various points below) that describe the annotation guidelines and document the routines for deriving various views of the data from the distributed OntoNotes database are included in the documentation directory of the distribution.

1.1 Summary Description of the OntoNotes Project

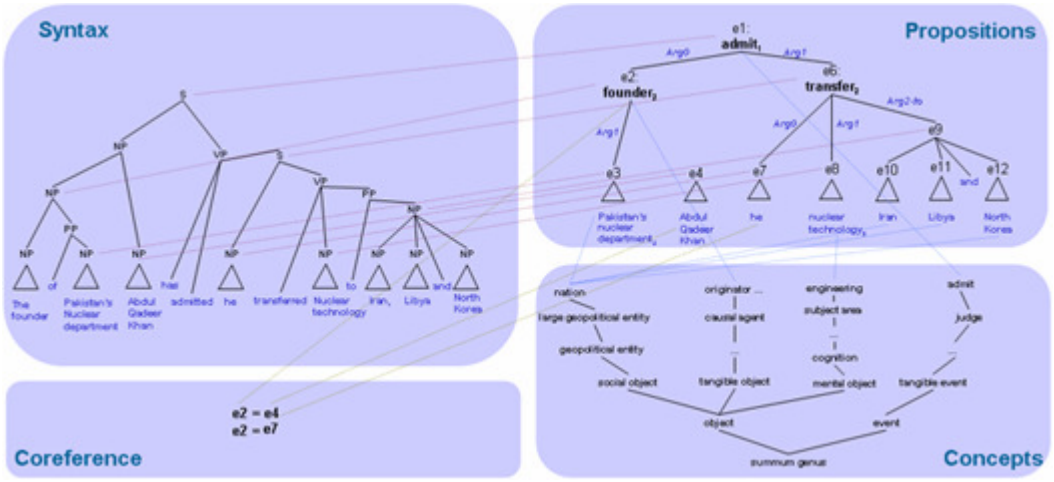
Natural language applications like machine translation, question answering, and summarization currently are forced to depend on impoverished text models like bags of words or n-grams, while the decisions that they are making ought to be based on the meanings of those words in context. That lack of semantics causes problems throughout the applications. Misinterpreting the meaning of an ambiguous word results in failing to extract data, incorrect alignments for translation, and ambiguous language models. Incorrect coreference resolution results in missed information (because a connection is not made) or incorrectly conflated information (due to false connections). Some richer semantic representation is badly needed.

The OntoNotes project is a collaborative effort between BBN Technologies, the University of Colorado, the University of Pennsylvania, and the University of Southern California's Information Sciences Institute to produce such a resource. It aims to annotate a large corpus comprising various genres of text (news, conversational telephone speech, weblogs, use net, broadcast, talk shows) in three languages (English, Chinese, and Arabic) with structural information (syntax and predicate argument structure) and shallow semantics (word sense linked to an ontology and coreference). OntoNotes builds on two time-tested resources, following the Penn Treebank for syntax and the Penn PropBank for predicate-argument structure. Its semantic representation will include word sense disambiguation for nouns and verbs, with each word sense connected to an ontology, and coreference. Over the course of the five-year program, our current goals call for annotation of over a million words each of English and Chinese, and half a million words of Arabic. Some details are provided in (OntoNotes, 2006)

Our plan is to make this resource available to the natural language research community so that decoders for these phenomena can be trained to generate the same structure in new documents. Lessons learned over the years have shown that the quality of annotation is crucial if it is going to be used for training machine learning algorithms. Taking this cue, we ensure that each layer of annotation in OntoNotes will have at least 90% inter-annotator agreement. Our pilot studies have shown that predicate structure, word sense, ontology linking, and coreference can all be annotated rapidly and with better than 90% consistency.

This level of semantic representation goes far beyond the entity and relation types currently targeted in the ACE program, since every concept in the text will be indexed, not just 100 pre-specified types. For example, consider this sentence: “The founder of Pakistan's nuclear program, Abdul Qadeer Khan, has admitted that he transferred nuclear

technology to Iran, Libya, and North Korea". In addition to the names, each of the nouns "founder", "program", and "technology" would be assigned a word sense and linked to an appropriate ontology node. The propositional connection signaled by "founder" between Khan and the program would also be marked. The verbs "admit" and "transfer" would have their word sense and argument structures identified and be linked to their equivalent ontology nodes. One argument of "admit" is "he", which would be connected by coreference to Khan, and the other is the entire transfer clause. The verb "transfer", in turn, has "he/Khan" as the agent, the technology as the item transferred, and the three nations Iran, Libya, and North Korea as the destination of the transfer. A graphical view of the representation is shown below:



Significant breakthroughs that change large sections of the field occur from time to time in Human Language Technology. The Penn Treebank in the late 1980s transformed parsing, and the statistical paradigm similarly transformed MT and other applications in the early 1990s. We believe that **OntoNotes** has the potential for being a breakthrough of this magnitude: it will be the first time ever that a semantic resource of this substantial size will be produced. As we have seen with the Treebank and WordNet, a publicly available resource unleashes an enormous amount of work internationally on algorithms and on the automated creation of semantic resources in numerous other domains and genres. We believe that this new level of semantic modeling will empower semantics-enabled applications to break the current accuracy barriers in transcription, translation, and question answering, fundamentally changing the nature of human language processing technology.

2 Annotation Layers

2.1 Treebank

The first level of OntoNotes analysis captures the syntactic structure of the text, following the approach taken in the Penn Treebank.

The Penn Treebank project, which began in 1989, has produced over three million words of skeletally parsed text from various genres. Among many other uses, the one million word corpus of English Wall Street Journal text included in Treebank-2 has fueled widespread and productive research efforts to improve the performance of statistical parsing engines. Treebanking efforts following the same general approach have also more recently been applied to other languages, including Chinese and Arabic.

While statistical parsers have often been evaluated on a reduced version of the Penn Treebank's structure, the OntoNotes goal of capturing literal semantics provides exactly the kind of context for which the full version of Treebank was initially designed. The function tags and trace information that are part of a full Treebank analysis will provide a crucial first step toward the OntoNotes analysis.

Within the OntoNotes project, the University of Pennsylvania will be providing the Treebank annotation for new genres of English text, and also contributing towards improving statistical parsing technology. The University of Colorado and the Linguistic Data Consortium will also be contributing Treebank data in Chinese and Arabic.

The Chinese Treebank (http://verbs.colorado.edu/chinese/ctb.html) is being developed at the University of Colorado, under the supervision of Prof. Martha Palmer and Nianwen Xue. The English Treebank (http://www.cis.upenn.edu/~treebank/) is being developed at the University of Pennsylvania under the supervision of Prof. Mitchell Marcus

2.2 PropBank

The propositional level of analysis is layered on top of the parse trees and identifies predicate constituents and their arguments in OntoNotes. This level of analysis is supplied by PropBank which is described below:

Robust syntactic parsers, made possible by new statistical techniques (Ratnaparkhi, 1997; Collins, 1998; Collins, 2000; Bangalore and Joshi, 1999; Charniak, 2000) and by the availability of large, hand-annotated training corpora (Marcus, Santorini, and Marcinkiewicz, 1993; Abeille, 2003), have had a major impact on the field of natural language processing in recent years. However, the syntactic analyses produced by these parsers are a long way from representing the full meaning of the sentence. As a simple example, in the sentences:

- John broke the window.
- The window broke.

A syntactic analysis will represent the window as the verb's direct object in the first sentence and its subject in the second, but does not indicate that it plays the same underlying semantic role in both cases. Note that both sentences are in the active voice,

and that this alternation between transitive and intransitive uses of the verb does not always occur, for example, in the sentences:

- The sergeant played taps.
- The sergeant played.

The subject has the same semantic role in both uses. The same verb can also undergo syntactic alternation, as in:

- Taps played quietly in the background.

and even in transitive uses, the role of the verb's direct object can differ:

- The sergeant played taps.
- The sergeant played a beat-up old bugle.

Alternation in the syntactic realization of semantic arguments is widespread, affecting most English verbs in some way, and the patterns exhibited by specific verbs vary widely (Levin, 1993). The syntactic annotation of the Penn Treebank makes it possible to identify the subjects and objects of verbs in sentences such as the above examples. While the Treebank provides semantic function tags such as temporal and locative for certain constituents (generally syntactic adjuncts), it does not distinguish the different roles played by a verb's grammatical subject or object in the above examples. Because the same verb used with the same syntactic subcategorization can assign different semantic roles, roles cannot be deterministically added to the Treebank by an automatic conversion process with 100% accuracy. Our semantic role annotation process begins with a rule-based automatic tagger, the output of which is then hand-corrected (see Section 4 for details).

The Proposition Bank aims to provide a broad-coverage hand annotated corpus of such phenomena, enabling the development of better domain-independent language understanding systems, and the quantitative study of how and why these syntactic alternations take place. We define a set of underlying semantic roles for each verb, and annotate each occurrence in the text of the original Penn Treebank. Each verb's roles are numbered, as in the following occurrences of the verb offer from our data:

- ...[$_{\text{Arg0}}$ the company] to ... offer [$_{\text{Arg1}}$ a 15% to 20% stake] [$_{\text{Arg2}}$ to the public]. (wsj 0345)
- ... [$_{\text{Arg0}}$ Sotheby's] ... offered [$_{\text{Arg2}}$ the Dorrance heirs] [$_{\text{Arg1}}$ a money-back guarantee] (wsj 1928)
- ... [$_{\text{Arg1}}$ an amendment] offered [$_{\text{Arg0}}$ by Rep. Peter DeFazio] ... (wsj 0107)
- ... [$_{\text{Arg2}}$ Subcontractors] will be offered [$_{\text{Arg1}}$ a settlement] ... (wsj 0187)

We believe that providing this level of semantic representation is important for applications including information extraction, question answering, and machine translation. Over the past decade, most work in the field of information extraction has shifted from complex rule-based systems designed to handle a wide variety of semantic phenomena including quantification, anaphora, aspect and modality (e.g. Alshawi, 1992), to more robust finite-state or statistical systems (Hobbs et al., 1997; Miller et al., 2000).

These newer systems rely on a shallower level of semantic representation, similar to the level we adopt for the Proposition Bank, but have also tended to be very domain specific. The systems are trained and evaluated on corpora annotated for semantic relations pertaining to, for example, corporate acquisitions or terrorist events. The Proposition Bank (PropBank) takes a similar approach in that we annotate predicates' semantic roles, while steering clear of the issues involved in quantification and discourse-level structure. By annotating semantic roles for every verb in our corpus, we provide a more domain-independent resource, which we hope will lead to more robust and broad-coverage natural language understanding systems.

The Proposition Bank focuses on the argument structure of verbs, and provides a complete corpus annotated with semantic roles, including roles traditionally viewed as arguments and as adjuncts. The Proposition Bank allows us for the first time to determine the frequency of syntactic variations in practice, the problems they pose for natural language understanding, and the strategies to which they may be susceptible.

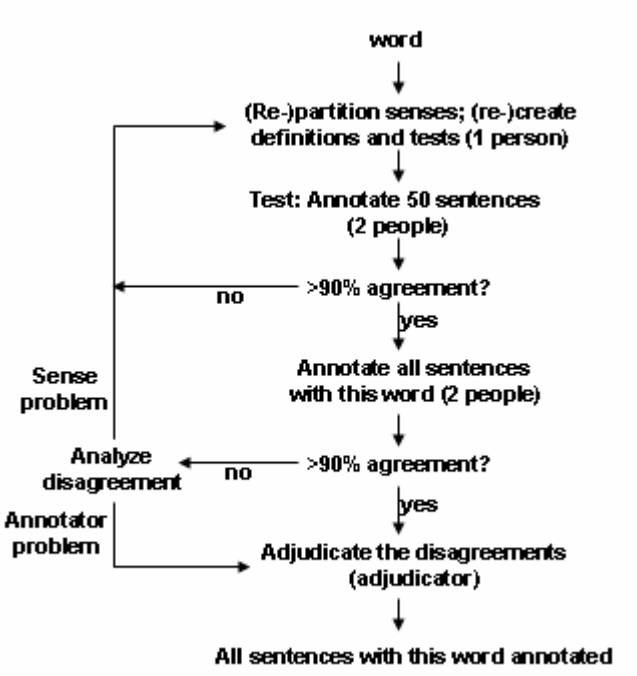
The Chinese PropBank (http://verbs.colorado.edu/chinese/cpb) is being developed at the University of Colorado, under the supervision of Prof. Martha Palmer and Nianwen Xue. The English PropBank (http://verbs.colorado.edu/mpalmer/palmer/projects/ace.html) is being developed at the University of Colorado and University of Pennsylvania, under the supervision of Prof. Martha Palmer and Olga Babko-Malaya.

2.3 Word Sense Annotation

Word sense ambiguity is a continuing major obstacle to accurate information extraction, summarization and machine translation. While WordNet has been an important resource in this area, the subtle fine-grained sense distinctions in it have not lent themselves to high agreement between human annotators or high automatic tagging performance. Building on results in grouping fine-grained WordNet senses into more coarse-grained senses that led to improved inter-annotator agreement (ITA) and system performance (Palmer et al., 2004; Palmer et al., 2006), we have developed a process for rapid sense inventory creation and annotation that also provides critical links between the grouped word senses and the Omega ontology (Philpot et al., 2005).

This process is based on recognizing that sense distinctions can be represented by linguists in a hierarchical structure, similar to a decision tree, that is rooted in very coarse-grained distinctions which become increasingly fine-grained until reaching WordNet (or similar) senses at the leaves. Sets of senses under specific nodes of the tree are grouped together into single entries, along with the syntactic and semantic criteria for their groupings, to be presented to the annotators.

As shown in Figure 1, a 50-sentence sample of instances is annotated and immediately checked for inter-annotator agreement. ITA scores below 90% lead to a revision and clarification of the groupings by the linguist. It is only after the groupings have passed the ITA hurdle that each individual group is combined with others with the same meaning and specified as a conceptual node in the ontology. In addition to higher accuracy, we find at least a three-fold increase in annotator productivity.



The word sense annotations for verbs is being carried out at the University of Colorado, under the supervision of Prof. Martha Palmer, and the same for nouns is being carried out at Information Sciences Institute, under the supervision of Prof. Eduard Hovy.

2.3.1 Verbs

Subcategorization frames and semantic classes of arguments play major roles in determining the groupings for verbs, as illustrated by the grouping for the 22 WN 2.1 senses for drive in Table 1. In addition to improved annotator productivity and accuracy, we predict a corresponding improvement in system performance. Training on this new data, Chen et al (2006) report 86.7% accuracy for verbs using a smoothed maximum entropy model and rich linguistic features. They also report state-of-the-art performance on fine-grained senses, but the results are more than 16% lower.

	WN1: “Can you drive a truck?”
	WN2: “drive to school”
GI: operating or traveling via a vehicle	WN3: “drive her to school”
	WN12: “this truck drives well”
NP (Agent) drive NP, NP drive PP	WN13: “he drives a taxi”
	WN14: “the car drove around the corner”
	WN:16: “drive the turnpike to work”
	WN4: “he drives me mad”
	WN6: “drive back the invaders”
G2: force to a position or stance	WN7: “she finally drove him to change
NP drive NP/PP/infinitival	jobs”
	WN8: “drive a nail”
	WN15: “drive the herd”

	WN22: “drive the game”
G3: to exert energy on behalf of something NP drive NP/infinitival	WN5: “her passion drives her” WN10: “he is driving away at his thesis”
G4: cause object to move rapidly by striking it NP drive NP	WN9: “drive the ball into the outfield” WN17 “drive a golf ball” WN18 “drive a ball”
G5: a directed course of conversation	WN11: “What are you driving at?”
G6: excavate horizontally, as in mining	WN19: “drive a tunnel through the mountain”
G7: cause to function or operate	WN20: “steam drives the engine”

2.3.2 Nouns

We follow a similar procedure for the annotation of nouns. The same individual who groups WordNet verb senses also creates noun senses, starting with WordNet and other dictionaries.

Certain nouns carry predicate structure; these include nominalizations (whose structure obviously is derived from their verbal form) and various types of relational nouns (like *father*, *President*, and *believer*, that express relations between entities, often stated using *of*). We have identified a limited set of these whose structural relations can be semi-automatically annotated with high accuracy.

2.3.3 Nominalizations and Eventive Noun Senses

In this section we present the definitions and possible uses of noun senses with the special designations *nominalization* and *eventive*. We have created lists of noun senses which are either nominalizations or eventives (or both), which are included in the Ontonotes word sense database. Noun senses on these lists largely correspond to noun senses in the sense definition files that include a nominalization or eventive feature, however, the lists are more restrictive and adhere to the criteria and definitions given below more rigorously.

Nominalizations have been identified so that the argument structures that they license can be correctely associated with elements of a nominal clause in which the nominalization appears. For example, in the sentence:

Achilles’ killing of Hector foreshadows the fall of Troy.

the nominal clause based on *killing* is *Achilles’ killing of Hector*. The NP *Achilles* is associated with arg0 and the NP *Hector* is associated with arg1. Although the nominalization senses have been identified, in the current release the arguments have not

yet been associated with the appropriate syntactic constituents; this will be done in a future version of Ontonotes..

The rationale for identifying some noun senses as *eventives* is somewhat different than it is for *nominalizations*. Eventive nouns often are also nominalizations, but not always. If a noun sense is eventive, it has a strong implication of a change of state in the situation it refers to, as well as a distinct and bounded time-frame. For example, in the sentence:

> We've just had a major fire.

the word *fire* is eventive, although there may be other non-eventive senses that appear in other contexts. The implication of the eventive sense of *fire* is that there was a prior state, an event onset, a state change, and a resulting state. Other modifiers may bring some aspect of the whole event process into focus, or remove some aspect from focus, but the basic *aktionsart* of the relevant word sense of *fire* is a temporally bounded event that results in a state change. By giving some noun senses this special designation, a given application (e.g. distillation) may be able to benefit, for example by employing temporal and causal reasoning. If it is known that there has been a *fire* event, subsequent references in the same text to *$50 million in property damage* may be determined to be, or be closely related to, the result state of the *fire* event.

The definitions and criteria for both nominalizations and eventive noun senses are given in more detail and with more examples in the following subsections.

Nominalization Senses of Nouns

Although it is traditional to speak of *words* (specifically *nouns*) as nominalizations, given the goals of the project, we find it more precise and useful to speak of particular *senses* of nouns as being *nominalization senses*. For example, it is imprecise to speak of the *word building* as a nominalization since only one *sense* of the word *building* is a *nominalization sense.* While the sense of the word invoked in the following sentence:

The building was made mostly of concrete and glass.

is *not* a nominalization sense, the sense invoked in:

The building of the Golden Gate Bridge was overseen by Joseph Strauss.

is a nominalization sense. The criteria we apply for identifying a sense of a noun as a nominalization sense are as follows:

> **(1)** The noun must relate transparently to a verb, and typically displays one of a set of nominalizing morphemes such as *–ment* (*govern/government*) and *–ion* (*contribute/contribution*) (see list below for others), though there are also many zero-derived nouns, such as *kill,* the noun, derived from *kill* the verb.
>
> **(2)** The noun must be able to be used in a clausal noun phrase, with its core verbal arguments related by semantically empty or very "light" licensers, such as genitive markers (as in "The Roman's destruction of the city...") or with the verb's usual particle or prepositional satellites (as in "John's longing for fame and fortune…").

The majority of the morphemes referred to in (1) above (mostly segmental suffixes) are as follows:

-ment	V -> N	(*govern* vs. *government*)
-ing	V -> N	(*trade* vs. *trading*)
-(t/s)ion	V -> N	(*contribute* vs. *contribution*)
-age	V -> N	(e.g. *pack* vs. *package*)
-t	V -> N	(*complain* vs. *complaint*)
-ure	V -> N	(*fail* vs. *failure*)
-ence, ance	V -> N	(*perform* vs. *performance*)
-al	mixed	(*propose* vs. *proposal*)
-y	V -> N	(*recover* vs. *recovery*)
stop →[s]	V -> N	(*succeed* vs. *success*)
-ity, ty	V -> N	(*prosper* vs. *prosperity*)
phonological devoicing	+voice = V, -voice = N	(*relieve* vs. *relief*)
stress-shift	word-final=V, word-initial=N	(*rebél* vs. *rébel*)

Discussion and Examples As noted in (1), in the case of zero-derived noun-verb pairs in which the noun has a nominalization sense (as in "the platoon's capture of the enemy scout") this noun must be related to a verb. What is more, the relation should be sufficiently transparent to enable speakers to access knowledge about the argument structure of the related verb. For example, although the noun *device* is related to the verb *devise*, it is difficult for native speakers to use the noun *device* naturally with the arguments associated with the verb. Thus, the following sentence sounds odd:

??Joe's device of the plan worried Mary.

One needs the form *devising* in order to obtain a natural-sounding construction, as in:

Joe's devising of the plan worried Mary.

Therefore, we exclude this sense of *device* from our list of nominalizations, but would include the relevant sense of the noun *devising*.

For the most part, the words we have identified as nominalizations conform with the traditional (linguistic) understanding of what a nominalization is. However, the following qualifications should be stated explicitly:

(i) Although we recognize that nominalizations may be based on verbs as well as other parts of speech (such as adjectives, as in *divinity*), we have included only nominalizations based on verbs.

(ii) We have omitted all nouns related to verbs with the agentive -er/-or marker (e.g. baker and hunter), as well as the majority of those with agentive/actor -ist or -ant/-ent (e.g. antagonist and assistant). The vast majority of words with these suffixes that we have identified have been kept in a separate list. The rationale behind this lies in the intended use of the nominalization sense status, which is to facilitate association of semantic arguments in nominal clauses with the syntactic elements within those nominal clauses. Since these agentive "nominalization" senses do not usually serve as a noun clause head, there will be no argument linking to facilitate.

Lastly, we note that the set of nominalization senses is fuzzy. Numerous cases of nominalization senses are not clearly good or bad, as in "the army's equipment of the troops" or "the recession of the river to its normal level". These sound more natural as "the army's equipping of the troops", and "the receding of the river to its normal level" but are certainly NOT on a par with (i.e. are not as bad as) the use of *device* in the sentence discussed earlier, "John's device of the plan worried Mary".

Eventive Senses of Nouns

Just as is the case for nominalizations, our view is that it is not strictly accurate to speak of *eventive nouns*, but rather to speak of eventive noun *senses*. For example, the sense of *party* accessed in a sentence like:

John had a wild party last Friday

is eventive, but the sense accessed in a sentence like

John is a member of the Republican party

is not. Also just as for nominalization senses, the set of eventive noun senses is fuzzy. We give the following definitional criteria (1-2) and a diagnostic test (3) for determining if a given noun sense is eventive.

(1) Activity causing a change of state

A noun sense is *eventive* when it refers to a single unbroken activity or process, occurring during a specific (though perhaps unknown) time period, that effects a change in the world of the discourse.

(2) Reference to Activity proper

The noun must refer to the actual activity or process, not merely to the result of the activity process.

(3) The noun patterns with eventive predicates in the 'have' test

A lexico-syntactic diagnostic test can be applied to many nouns to determine if they are eventive, as described by the following heuristic (Belvin, 1993):

(i) Create as natural sounding a sentence as possible using the construction ***X had <NP>***, where <NP> is a noun phrase headed by the noun in question; for example if our noun is "party", we start with the sentence template "X had a party". Then:

(ii) Check if the sentence can be used in a *present progressive construction*, such as:

John is having a party.

If this sounds felicitous, it adds evidence to the noun being eventive. If it sounds odd, it adds evidence that the noun is stative.

(iii) Check if the sentence can be used in a pseudocleft construction, such as:

What John did was have a party.

If this sounds felicitous, it adds evidence to the noun being eventive. If it sounds odd, it adds evidence that the noun is stative.

(iv) Check if the sentence suggests iterative/habitual action using the simple present tense, such as:

?John has a party.

If so (as in this case, e.g., "John has a party every Friday"), it adds evidence that the noun is eventive. If the sentence suggests that the situation is taking place at the very moment that it is uttered, it adds evidence that the noun is stative (as for example in "John has a cold").

Discussion and Examples Notice that one of the criteria for being an eventive noun sense is that the noun does NOT have to be transparently related to a verb, and it does NOT have to license arguments in a clausal NP structure. Eventive noun senses frequently do show these characteristics, but it is not a requirement; this often distinguishes this noun sense type from nominalizations, for which these two criteria are required. However, there is a very significant *intersection* of eventive senses and nominalization senses.

Returning to the definitional criteria of eventive noun senses above, we briefly consider the characteristics of the "change within the world of discourse". The noun sense in question is *less* eventive to the extent that this change is not singular, homogeneous, or occurring over a short period, but instead is a collection of changes of different kinds, and possibly over a longer period of time. Where exactly an event ceases to be a change and becomes a gradually changing state is a matter of choice, depending on the timescale of the perspective being taken in the discourse. Thus "war" may be (weakly) eventive in the phrase "WW II" , if it is seen as a point 'event' within the span of a century of more, whereas it is very unlikely to be so in "the 100-years' War" over the same time span. Similarly, the weathering of the Sphinx over centuries is not a canonical event, even though it is a rather homogeneous and continuous process.

Additional evidence for a noun sense being eventive is: (i) the existence of a corresponding verb form; (ii) the noun sense occurring with similar patterns of complements (their hope for peace, they hoped for peace); and (iii) the presence in the noun of a recognized nominalization suffix. However, as noted earlier not all nominalization senses are eventive (e.g. *an understanding of the issues...*) and not all eventive nouns are nominalizations (e.g. *party*).

To further clarify the intended meaning of the term *eventive nouns senses,* we here provide some examples of eventive and stative nouns illustrating aspects of the definition:

- "cake" in "he baked a cake" is clearly not eventive, being the result of some activity
- "auction" in "there was an auction last night" is eventive, despite consisting of several smaller events—the whole thing is contiguous and does effect a change, in the world, namely the change(s) of ownership
- "trouble" in "don't go to a lot of trouble with John's dinner tonight", and "I had some trouble with my car today" is eventive
- "attitude" in "he assumed a convincing attitude of a despotic king in the school play" is not eventive since the attitude is the *result* of the assumption of a stance and is therefore a state
- "record" in "his record is impressive" is not eventive since it is merely the record of the change
- seasonal or weather nouns such as "spring", "winter", "freeze", "drought" can be eventive depending on the time scale involved relative to the current (typical, default) perspective scale. Thus in "the freeze of Dec 15, 1903 was the worst of a decade" is eventive, being one night in ten years and with a clear change of state entailed.

2.4 Ontology

The Omega ontology (in particular, Omega 5) provides the semantic framework for the OntoNotes annotation. Word senses in OntoNotes are pooled into groups with (near-) identical meanings (similar to synsets in Wordnet), and these pools, treated as concepts, become ontology nodes in Omega. Each pool will ultimately be linked into Omega, allowing its parent and sibling modes to provide semantic generalizations of the concept conveyed by the word(s) whose senses are contained in the pool. Ultimately, the pools also furnish a place to store additional axioms to help in interpreting the entities and relations conveyed.

At this time, enough related word senses have been defined and annotated to begin the ontologizing process. In this Year 2 release, therefore, we provide the first version of Omega 5. While it is not very large yet, the infrastructure and all necessary parts are in place, and pools will be linked into Omega 5 throughout Year 3.

Omega 5 consists of two parts: an Upper Model of approximately 200 nodes and the ontology body. Each Ontology node represents a conceptualization. Upper Model nodes are hand-build to represent high-level important generalizations that help organize the remaining nodes. The Upper Model is currently organized into two primary branches: Objects and Eventualities. (In later versions, Omega will also contain a branch for Qualities/Properties). The Object nodes taxonomize all objects/entities (typically, pools of noun senses) into approximately 35 classes, and the Eventuality nodes define approximately 20 classes for processes/events (typically, pools of verb senses). Upper Model nodes introduce definitional features— atomic terms like +*concrete*, -*concrete*, +*animate*, etc.—that specify aspects of the concepts they govern.

Regarding the ontology body, nodes are formed out of OntoNotes senses as follows. Each sense of a word in the OntoNotes corpus is combined (pooled) with senses of other OntoNotes words that carry the same meaning and is verified independently by two or more 'sense poolers'. The verification process is described in (Yu et al. 2007). Also associated with each pool can be one or more additional features—atomic terms like the features of the Upper Model—that specify some aspects of the concept, and help differentiate it from its nearly similar pools. At time of writing, over 4000 features have been defined, but are not yet finalized or complete.

Each pool of senses, together with all definitions, examples, and annotated sentence instances, forms a concept in Omega. These concepts are attached to the Omega Upper Model at appropriate concepts (almost always, the leaves of the Upper Model). They therefore inherit the features pertinent for their attachment point concept. For some pools, subordination links have been defined, which provide a very shallow amount of hierarchicalization of the ontology body. However, in general there is no plan to produce a fixed, deep, and predefined hierarchicalization of Omega concepts below the Upper Model. The user can impose a hierarchical structure at any time by specifying the desired sequence of importance of concept features. (For example, by specifying first *age* and then *gender*, the substructure under Person is first Adult and Child and then below that Man, Woman, Boy, Girl. In contrast, specifying first *gender* and then *age* gives the substructure MalePerson and FemalePerson and then Man, Boy, Woman, Girl.)

This release includes about 1000 Object pool concepts and 50 Eventuality sense pool concepts that are subordinated to the Upper Model, and a few hundred more Eventualities that are not yet subordinated.

At present, Omega 4 (the previous version of the ontology) is a 120,000-node terminological ontology constructed at USC ISI as the synthesis of WordNet 2.0 (Miller 1990; Fellbaum 1998), a lexically oriented network constructed on general cognitive principles, and Mikrokosmos (Mahesh 1996; O'Hara et al. 1998), a conceptual resource originally conceived to support translation, whose result is subordinated under a new upper model, created expressly in order to facilitate the merging of lower models into a functional whole. Omega, like its close predecessor SENSUS (Knight and Luk 1994), can be characterized as a shallow, lexically oriented, term taxonomy; by far the majority of its concepts can be stated in English using a single word. Omega contains no formal concept definitions and only relatively few interconnections (semantic relations) between concepts. By making few commitments to any specific theories of semantics or particular representations, Omega enjoys a malleability that has allowed it to be used in a variety of applications, including question answering and information integration. A major aim in constructing Omega was to leverage the strengths and minimize the weaknesses of the two major constituents: to have a large, lexically rich resource work with a clear comprehensive organization, supporting both inference and lexical access.

The Omega (http://omega.isi.edu/) ontology (Philpot et al., 2005) is being developed at the Information Sciences Institute under the supervision of Prof. Eduard Hovy

2.5 Coreference

The coreference annotation project is being carried out at BBN Technologies under the supervision of Ralph Weischedel and Lance Ramshaw

The goal of OntoNotes coreference annotation and modeling is to fill in the coreference portion of the shallow semantic understanding of the text that OntoNotes is targeting. For example, in "She had a good suggestion and it was unanimously accepted", we mark a case of IDENT coreference (identical reference) between "a good suggestion" and "it", which then allows correct interpretation of the subject argument of the "accepted" predicate.

Names, nominal mentions, and pronouns can be marked as coreferent. Verbs that are coreferenced with a noun phrase can also be marked as IDENT; for example "grew" and "the strong growth" would be linked in the following case: "Sales of passenger cars grew 22%. The strong growth followed year-to-year increases." In addition, in 'pro-drop' languages like Chinese, coreference annotation can be applied to a "*pro*" element taken from the Treebank parse which serves as a placeholder for the missing pronoun.

In order to keep the annotation feasible at high agreement levels, only intra-document anaphoric coreference is being marked. Furthermore, while annotation is not limited to any fixed list of target entity types, noun phrases that are generic, underspecified, or abstract are not annotated.

Attributive NPs are not annotated as coreference because the meaning in such cases can be more appropriately taken from other elements in the text. For example, in "New York is a large city", the connection between New York and the attributive NP "a large city" comes from the meaning of the copula "is". Similarly, in "Mary calls New York heaven", the connection comes from the meaning of the verb "call". Thus these cases are not marked as IDENT coreference.

Appositive constructions are marked with special labels. For example, in "Washington, the capital city, is on the East coast", we annotate an appositive link between Washington (marked as HEAD) and "the capital city" (marked as ATTRIBUTE). The intended semantic connection can then be filled in by supplying the implicit copula.

2.6 Entity Names Annotation

Names (often referred to as "Named Entities") are annotated according to the following set of types:

PERSON	People, including fictional
NORP	Nationalities or religious or political groups
FACILITY	Buildings, airports, highways, bridges, etc.
ORGANIZATION	Companies, agencies, institutions, etc.
GPE	Countries, cities, states
LOCATION	Non-GPE locations, mountain ranges, bodies of water
PRODUCT	Vehicles, weapons, foods, etc. (Not services)

EVENT	Named hurricanes, battles, wars, sports events, etc.
WORK OF ART	Titles of books, songs, etc.
LAW	Named documents made into laws
LANGUAGE	Any named language

The following values are also annotated similarly to names:

DATE	Absolute or relative dates or periods
TIME	Times smaller than a day
PERCENT	Percentage (including "%")
MONEY	Monetary values, including unit
QUANTITY	Measurements, as of weight or distance
ORDINAL	"first", "second"
CARDINAL	Numerals that do not fall under another type

3 Corpus Plan

The OntoNotes project has laid out a five-year plan to achieve substantial coverage in various genres and in all three GALE languages.

Full OntoNotes Annotation.	end Y1	end Y2	end Y3	end Y4	end Y5
English					
Newswire	300K		250K ECTB		
Broadcast News		200K			
Broadcast Conversation			200K		
Newsgroups				200K	
Weblogs				200K	
Conversational Telephone Speech					100K, 100K*
Chinese					
Newswire	250K ECTB			150K	
Broadcast News		300K			
Broadcast Conversation			150K		
Newsgroups				150K*	
Weblogs					100K, 50K*
Arabic					
Newswire		100K	100K	100K	
Broadcast News				100K	100K
Total	550K	600K	700K	750K, 150K*	300K, 150K*

(* Wordsense annotation only)

The current Year 2 release covers newswire and broadcast news data in English and Chinese. The newswire portion includes 300K of English Wall St. Journal newswire and 250K of Chinese newswire (100K of Xinhua newswire and 150K of Sinorama news magazine text). The broadcast news portion includes 200K of English data and 300K of Chinese data, both taken from the TDT-4 collection.

While some Arabic OntoNotes annotation has been completed, the parse trees for this corpus are currently being revised by the LDC. When the revised trees are available, the OntoNotes annotation will need to be updated to be consistent with them before it can be incorporated into our OntoNotes database format.

Because of our staged approach to annotation, Treebank coverage of corpora is often available well in advance of full OntoNotes annotation. For the English and Chinese broadcast conversation corpora that will be released at the end of Year 3, the Treebank parses have already been made available to the other GALE participants. This data includes 100K of parallel data, 50K of which was originally English, translated into Chinese, and 50K of which was originally Chinese, translated into English.

4 English Release Notes

4.1 English Year 1 and 2 Corpora

The English OntoNotes corpus so far includes 300K of newswire from Year 1 and 200K of broadcast news from Year 2.

The newswire corpus from Year 1 is a 300K portion of the Penn Treebank 2 Wall Street Journal corpus. Documents were selected so as to try to avoid stories that were strictly financial such as daily market reports.

The broadcast news data is a 200K portion selected from the TDT4 corpus, using documents that had previously been annotated by the LDC as part of the ACE (Automatic Content Extraction) program.

4.2 Treebank Notes

The annotation of syntactic structure trees in our English newswire data is taken with few changes straight from Penn Treebank 2. The syntactic structure for the broadcast news data was annotated from scratch as part of this project. The accompanying documentation directory includes the following documents that describe the guidelines used in this annotation:

- english-treebank-postags.ps: Part of Speech tagging guidelines.
- english-treebank2-bracketing.ps: Syntactic structure guidelines for Treebank 2.
- english-treebank-guidelines-addendum.pdf: Modifications in the syntactic structure guidelines since Treebank 2.

A number of revisions in the tree structures that were made to align them more closely with the PropBank annotation are also described further in Section 4.4 below.

4.3 PropBank Notes

The PropBank annotation of propositions and arguments in our English newswire corpus is largely taken from the previously released "PropBank 1", though some changes were made to align the propositions more closely with the Treebank annotation. The PropBank annotation for the broadcast news data was done as part of this project.

In the WSJ 300K corpus, 33,147 propositions are annotated, covering all verbs (with the exception of auxiliaries and the verb "be") and some eventive nouns. (Coverage of the verb "be" for this corpus will be added during Year 2.) The total number of verb types annotated is 2428, which have 3120 total framesets. 1,733 noun instances are annotated, with 73 noun types. The YR2 release involves the 200K English Broadcast News corpus, consisting of 33,800 propositions for 1,937 verb types which were double-annotated and adjudicated, including the "be" verb instances. 20 new Frame Files were added. Each annotation includes a link to the relevant frameset entry. For a detailed description of the PropBank data fields and formats, see Section 6.5 below. The annotation guidelines are included in the documentation directory:

- English-propbank.doc: English propbank annotation guidelines

4.4 Treebank/Propbank Merge Notes

In Propbank 1.0, propbank annotators often made choices that do not conform with the Treebank parses. The discrepancies between the two sources obstruct the study of the syntax and semantic interfaces and pose immediate problems to an automatic semantic role labeling system. Some changes were necessary in both the Treebank and PropBank as part of OntoNotes to address this issue. More details about the Treebank/ProbBank discrepancies and their reconciliation can be found in Babko-Malaya et al (2006), which can also be found in the file "treebank-propbank-merge.pdf" in the documentation directory of this distribution.

4.4.1 Treebank Changes

The changes that were made on the Treebank side to help enable the Treebank/PropBank merge included a reorganization of verbal complementation and control that distinguished subject control from raising, a redrawing of the boundary between verbs that take small clauses and those that take secondary predicates, and a revised treatment of parentheticals, among others. A more detailed description of these changes can be found in the file "treebank-propbank-merge-treebank-changes.pdf" in the documentation directory of this distribution.

Note that certain of these Treebank guideline changes turned out to be too costly to update in the existing Treebank data. In particular, the revised guidelines call for using NML ("nominal") constituents as additional substructure within NP premodifiers. While this has been done in the newly-parsed broadcast news data, that change has not yet been made in the parse trees for the 300K of newswire data.

4.4.2 Propbank changes

After the changes are made to the Treebank, the Propbank annotation was realigned with the Treebank. Mostly this involves shifting the Propbank pointers to match the appropriate constituents in the parse tree. As a result of the Treebank changes with regard to the small clauses, the frame files for certain verbs that typically take small clauses as their complements were changed as well. There are also stylistic changes with regard to how the trace chains are represented in the Propbank. In particular, in the previous version of the propbank, the head of a relative clause is chained together with the relative pronoun as well as the trace that is coindexed with the relative pronoun in the Treebank. This chain as a whole was assigned a semantic role label. In the current release of the revised propbank, the trace is only chained to the relative pronoun and they are assigned an argument label. The semantic relation between the relative pronoun and the head of the relative clause is annotated as a separate link, LINK-SLC (for SeLectional Constraint link). The second stylistic change is that certain PROs in the Treebank are now annotated as LINK-PCR, for Pragmatic Coreference Link.

4.5 Word Sense Notes

In this Year 1 English release, we targeted substantial but not complete word sense coverage for polysemous nouns and verbs in the 300K WSJ corpus. For verbs, 57,626 instances of 737 of the most frequent polysemous cases have been annotated, and for nouns, the instances of 1092 of the most frequent polysemous cases have been covered. (A small number of very frequent verbs including the verb "be" proved problematic in terms of defining the sense inventory; coverage for these was added to the Year 2 corpus.) The YR 2 corpus includes the annotation of the same verbs on the new data (200k Broadcast News and 250K ECTB) as well as annotation of additional 700 verbs. These verbs have all been sample annotated, but the YR 2 data release only includes double annotated adjudicated data for 244 of the verbs. In future years of the OntoNotes project, additional words will be annotated, increasing the density of coverage of the materials included in this initial release. Around 500 verb senses have also been "pooled" into around 400 pools (see Section 2.4). These cover "communication" verbs and "motion" verbs.

For annotated words, an OntoNotes word sense number is listed in the database for each instance of the word. The accompanying sense inventory file documents the intended meaning of each numbered OntoNotes sense.

Some of the English word sense annotation has not yet been fully double annotated and adjudicated. Single-annotated word senses can be distinguished in the data on the basis of an "adjudicated" flag stored in the DB record for each word.

4.6 Coreference Notes

The guidelines for our English coreference annotation can be found in the file "english-coref.pdf" in the accompanying documentation directory.

4.7 Name Annotation Notes

The name annotation of the English data follows the 11 entity name types and 7 value types described in Section 2.6.

5 Chinese Release Notes

5.1 Chinese Year 1and 2 Corpora

The Chinese portion of OntoNotes 2.0 includes 250K words of newswire data and 300K words of broadcast news.

The newswire data (the Year 1 Chinese corpus) is taken from the Chinese Treebank 5.0. That 250K includes 100K of Xinhua news data (chtb_001.fid to chtb_325.fid) and 150K of data from the Sinorama news magazine (chtb_1001.fid to chtb_1078.fid).

The broadcast news data (the Chinese corpus for Year 2) is 274K words taken from TDT4, and selected from data that was annotated by the LDC for the Automatic Content Extraction (ACE) program. These files have been assigned numbers chtb_2000.fid to chtb_3145.fid.

5.2 Treebank Notes

The annotation of syntactic structure trees for our Year 1 Chinese newswire data was taken from the Chinese Treebank 5.0 and updated with some corrections. Some of known problems, like multiple tree nodes at the top level, were fixed. We also fixed some inconsistent annotations for object control verbs. The residual Traditional Chinese characters in the Sinorama portion of the data, the result of incomplete automatic conversion, have been manually normalized to Simplified Chinese characters.

The syntactic structure annotation for the Chinese Year 2 corpus was done entirely under the GALE OntoNotes program.

The accompanying documentation directory includes the following documents that describe the guidelines used in this annotation. More detailed description about the Chinese Treebank can also be found in Xue et al (2005).

- chinese-treebank-postags.pdf: Part of Speech tagging guidelines for the Chinese Treebank
- chinese-treebank-segmentation.pdf: Word segmentation guidelines for the Chinese Treebank
- chinese-treebank-parses.pdf: Syntactic structure guidelines for the Chinese Treebank.
- chinese-treebank-parses-bn-addendum.pdf: Addendum for the broadcast news portion of the data that has noises from the transcription of the spoken language.

The content used in CTB 5.0 comes from the following newswire sources:

698 articles Xinhua (1994-1998)

55 articles Information Services Department of HKSAR (1997)

132 articles Sinorama magazine, Taiwan (1996-1998 & 2000-2001)

5.3 PropBank Notes

For the Chinese newswire data, the annotation of the verbs in the Xinhua news portion of the data is taken from Chinese Proposition Bank 1.0, which has already been released through the LDC, but the annotation of the predicate-argument structure of the nouns, which are primarily nominalizations, has not been previously released. The Sinorama portion of the data, both for verbs and nouns, has not been previously released. The Year 1 release of the Chinese Propbank for the newswire data has annotation for 7,745 verb types in 41,327 propositions and 995 noun types in 7,036 propositions.

For the broadcast news data, the Year 2 release contains 46,241 propositions for 4,335 verbs. Unfortunately, full double annotation and adjudication for the Year 2 broadcast news data has not yet been completed, so the PropBank annotation for this data is currently only single-annotated.

The accompanying documentation directory contains the annotation guidelines for the Chinese Proposition Bank:

- chinese-propbank.pdf: annotation guidelines for the Chinese Proposition Bank

This release also contains the *frame files* for each verb or noun annotated in this corpus, which specify the argument structure (semantic roles) for each predicate. The frame files are effectively lexical guidelines for the propbank annotation. The semantic roles annotated in this data can only be interpreted with respect to these frame files. Detailed descriptions of the Chinese Proposition Bank can be found in an article by Xue and Palmer, currently under review for *Natural Language Engineering*.

5.4 Word Sense Notes

In our Year 1 Chinese release, we targeted substantial but not complete word sense coverage for polysemous nouns and verbs. For verbs, all 24,727 instances of about 300 of the most frequent polysemous verbs were annotated. For nouns, about 216 of the most frequent polysemous nouns were annotated. In our Year 2 Release, we have included 25,149 verb instances of those same 300 verbs from Year 1 as found in the new broadcast news data. We have also annotated new verbs in both the newswire and broadcast news data, selected from those verbs not yet covered ones that occur most frequently in the new broadcast news corpus. Similarly, for nouns, we have annotated the nouns from last year in the new data, and added coverage of additional nouns in both the newswire and broadcast news data. As the OntoNotes project continues, additional words will continue to be annotated, increasing the density of coverage of the materials included in this initial release.

For annotated words, an OntoNotes word sense number is listed in the database for each instance of the word. The accompanying sense inventory file documents the intended meaning of each numbered OntoNotes sense.

Some of the Chinese word sense annotation has not yet been fully double annotated and adjudicated. Single-annotated word senses can be distinguished in the data using the value to the “adjudicated” flag in the DB record for the word.

5.5 Coreference Notes

The guidelines for our Chinese coreference annotation can be found in the file "chinese-coref.pdf" in the accompanying documentation directory.

Coreference coverage of the broadcast news portion is not yet complete. Because even single annotation data could be useful, it has been included in the release. There is an "adjudicated" flag in the DB record for each file, which can be used to separate out fully double-annotated and adjudicated files from those for which only single annotation is available.

5.6 Name Annotation Notes

The name annotation of the Chinese data follows the 11 entity name types and 7 value types described in Section 2.6.

Unfortunately, the name annotation for the Chinese data is not yet complete. The Sinorama portion of the Year 1 newswire corpus and certain files in the Year 2 broadcast news corpus do not yet have name annotation done. Such files can be recognized by the absence of any name annotation.

6 Database, Views, Supplementary Data, and Data Access Guide

This section describes the integrated database in which all of the OntoNotes annotation is stored, and various ways of accessing the data.

Functions are provided that can output various "views", text files that encode a single layer of annotation, usually in a format very similar to that produced by the original annotation tools. There is also an "OntoNotes Normal Form" view, which combines all of the levels in a single readable version.

As an alternative method for inspecting the data, the separate OntoViewer utility, which is included with this release, provides a flexible interactive view of the different annotation layers, including a view that displays all of the nested propositions conveyed by each sentence.

6.1 How the OntoNotes Data is Organized

The normative version of the OntoNotes annotation is a relational database, in which the various layers of annotation for both the English and Chinese corpora are merged. It was created by loading the separate Treebank, PropBank, word sense, and coreference sources and merging them into a set of linked relational database tables. A dump-file image of the resulting database is included in this distribution (y2-ontonotes.sql), along with the original source files and the code that was used to do the merge.

The source files for each of the layers of annotation are included in the data directory of the distribution, using separate files for each layer of annotation of each corpus document file. The following filename extensions are used for each of the five layers:

- parse
- prop
- sense
- coref
- name

These are versions of the annotation files that have been output from the database to ensure consistency. The ontology information is in stored as one "upper-model.xml" file that represents the ontology upper model, and a collection of XML files, representing the sense pools, in the "sense-pools" subdirectory of the ontology directory. These are the source files representing the ontology, but in addition, we also have two .dot files which represent the same data in the Graphviz (www.graphviz.org) ".dot" file format – one for just the upper model and one for the entire ontology – upper model and sense pools combined.

In addition to the annotation-level views of the data that can be extracted from the database, there is also an "OntoNotes Normal Form" (ONF) view, which tries to render the merged annotation in human-readable form. The ONF are found in the distribution in their own ontonotes-normal-form directory.

The following subsections describe the database design, the different annotation views, and the OntoNotes Normal Form view. There is also a section describing the supplementary data files in which the PropBank propositional frames and the OntoNotes word senses are defined. Finally, a section provides pointers to the documentation for the scripts that have been used to do the merging of the different annotation layers and to generate the various views, since users may find those routines helpful for writing their own database queries or views, or for extending the schema.

6.2 OntoNotes Annotation Database

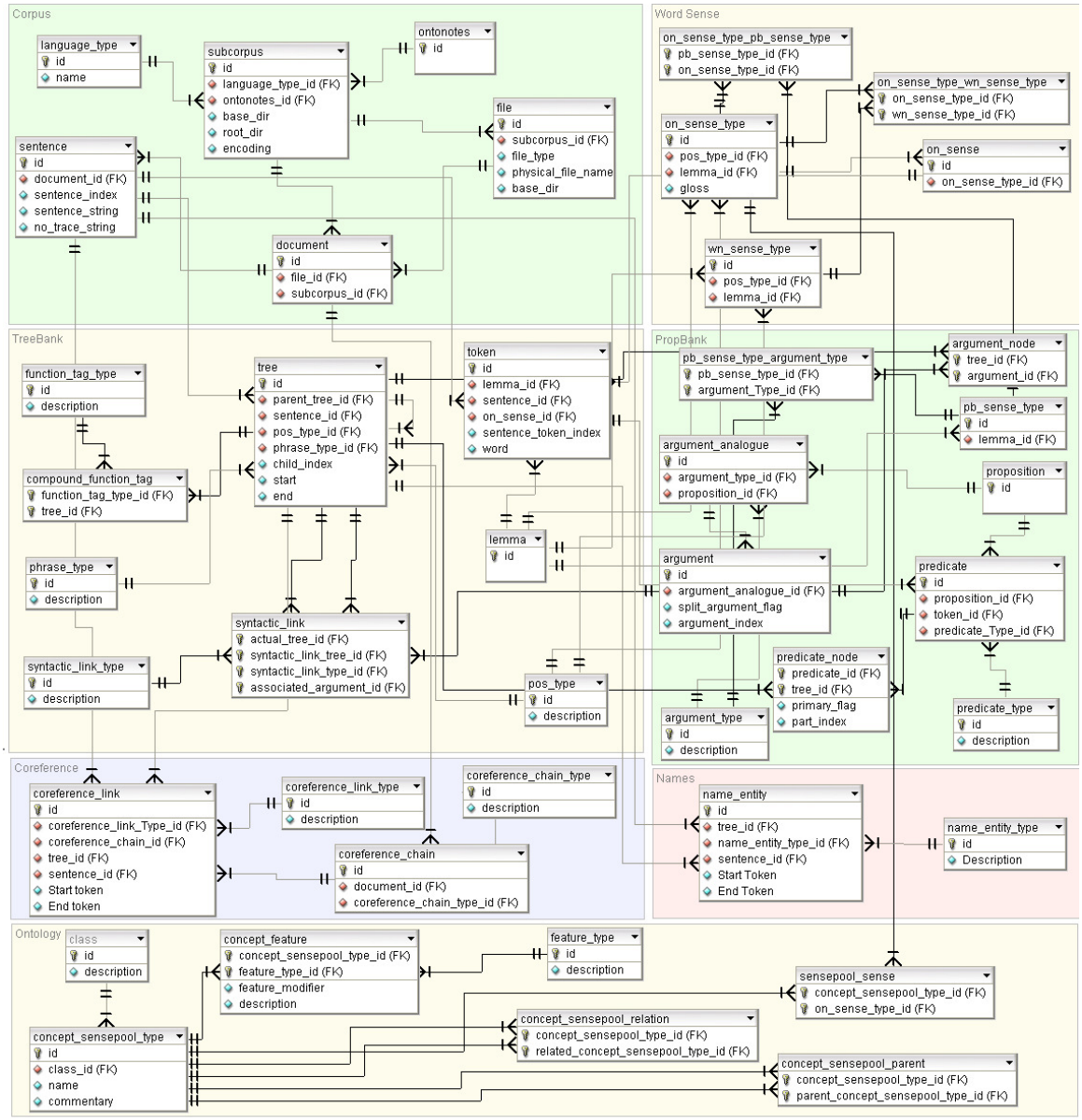


Figure 1: The OntoNotes Database Schema

The OntoNotes database schema is shown in Figure 1. The database tables are shown divided into six logical blocks, with one block for the textual corpus data, and then a

block for each type of semantic annotation: Treebank, Proposition Bank, Word Sense, Coreference, and Name Entities. Each of the annotation types involves adding additional meta information to the corpus. The basic units of annotations are the tokens as defined by the tokenization scheme in the Treebank; all of the annotation layers abide by this constraint. In addition, most of the text spans (with a few exceptions) are in alignment with the nodes of trees in the Treebank. The exceptional cases are addressed by using token start and end indices to define the spans.

The directory and file structure of the raw OntoNotes data organization is as shown in Figures 2 and 3 respectively. Since the smallest coherent piece is a document, we have created document-specific annotation files. The file extension specifies the annotation type.

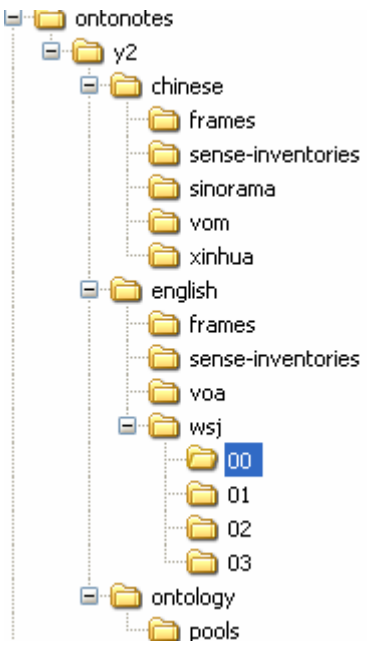


Figure 2: Directory structure for the raw OntoNotes Data

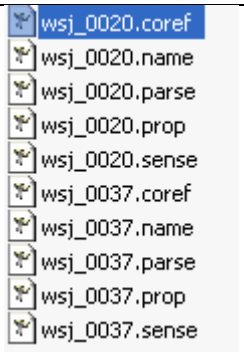


Figure 3: File structure of the raw OntoNotes data.

The database manipulation API that is provided with this release, and which is described in more detail in the "API Reference" accompanying this document, reads in this structure and populates the aforementioned database. As part of the API, we have provided mechanisms to produce the individual views as they are represented in the respective raw documents, as well as a more human-readable composite view. The former ensures that the data that it represents has been tested for consistency, since the database loading routines will not load any data that it finds to be inconsistent. The data is organized in folders which represent each text source. Further, following the convention of the Penn Treebank, we have created sections inside each source with a maximum of hundred files in each section. The mapping information of these new filenames to the originals are available in the map.txt files in the top-level "english" and "chinese" directories. There is no mapping information for the Xinhua and Sinorama

portions of the Chinese data, and WSJ portion of the English data, since these were already available in mapped form, and we did not re-map them. Furthermore, these views can also be useful for regenerating the raw data after any possible manipulations on the database. We will take a look at each such view in the following subsections.

6.3 OntoNotes Normal Form (ONF) View

The OntoNotes Normal Form (ONF) is a textual view that formats the combined layers of OntoNotes annotation for human review, including the text, the parse, the propositions, and the coreference chains.

Here is a sample of the OntoNotes Normal Form output for a sentence:

```
================================================================================
--------------------------------------------------------------------------------
id: 0@wsj/06/wsj_0655@wsj@en@on
--------------------------------------------------------------------------------

sentence: Nicaraguan President Daniel Ortega may have accomplished over
the weekend what his U.S. antagonists have failed *-1 to do *T*-2 :
*PRO* revive a constituency for the Contra rebels .
--------

parse tree:
---------

    (TOP (S (NP-SBJ (JJ Nicaraguan)
                    (NNP President)
                    (NNP Daniel)
                    (NNP Ortega))
            (VP (MD may)
                (VP (VB have)
                    (VP (VBN accomplished)
                        (PP-TMP (IN over)
                                (NP (DT the)
                                    (NN weekend)))
                        (NP (SBAR-NOM (WHNP-2 (WP what))
                                      (S (NP-SBJ-1 (PRP$ his)
                                                   (NNP U.S.)
                                                   (NNS antagonists))
                                         (VP (VBP have)
                                             (VP (VBN failed)
                                                 (S (NP-SBJ (-NONE- *-
1))
                                                    (VP (TO to)
                                                        (VP (VB do)
                                                            (NP (-NONE-
*T*-2)))))))))))
                            (: :)
                            (S (NP-SBJ (-NONE- *PRO*))
                               (VP (VB revive)
                                   (NP (NP (DT a)
                                           (NN constituency))
                                       (PP (IN for)
                                           (NP (DT the)
                                               (NNP Contra)
```

```
                                                    (NNS rebels))))))))))
            (. .))))

----

sense and proposition:
---------------------

Nicaraguan          0
President           1
Daniel              2
Ortega              3
may                 4
have                5
accomplished        6
--------------------------------------------------
predicate: accomplish;     pb_sense: 01

 ARG1        ->     10:3   -> what his U.S. antagonists have failed *-1
to do *T*-2 : *PRO* revive a constituency for the Contra rebels
 ARGM-MOD    ->     4:0    -> may
 ARG0        ->     0:1    -> Nicaraguan President Daniel Ortega
 ARGM-TMP    ->     7:1    -> over the weekend
--------------------------------------------------
over                7
the                 8
weekend             9
what                10
his                 11
U.S.                12
antagonists         13
have                14
failed              15 on_sense: 1
--------------------------------------------------
predicate: fail;     pb_sense: 01

 ARG1        ->     11:1   -> his U.S. antagonists
 ARG2        ->     19:0   -> *T*-2  ->     10:1   -> what
--------------------------------------------------
*-1                 16
to                  17
do                  18 on_sense: 1
--------------------------------------------------
predicate: do;     pb_sense: 02

 ARG0        ->     16:0   -> *-1  ->     11:1   -> his U.S.
antagonists
 ARG1        ->     19:0   -> *T*-2  ->     10:1   -> what
--------------------------------------------------
*T*-2               19
:                   20
*PRO*               21
revive              22 on_sense: 1
--------------------------------------------------
predicate: revive;     pb_sense: 01

 ARG1        ->     23:2   -> a constituency for the Contra rebels
```

```
 ARG0        ->     0:1    -> Nicaraguan President Daniel Ortega  ->
21:0   -> *PRO*
-------------------------------------------------
a                   23
constituency        24
for                 25
the                 26
Contra              27
rebels              28
.                   29

-------------------------------------------------
---------------- COREFERENCE CHAINS ----------------
-------------------------------------------------
DOCNO: wsj/06/wsj_0655@wsj@en@on

CHAIN: IDENT@70@wsj/06/wsj_0655@wsj@en@on
LINKS:
 IDENT@0:27:27@IDENT@70@wsj/06/wsj_0655@wsj@en@on: Contra
 IDENT@6:18:18@IDENT@70@wsj/06/wsj_0655@wsj@en@on: Contra
 IDENT@16:32:32@IDENT@70@wsj/06/wsj_0655@wsj@en@on: Contra
 IDENT@24:11:11@IDENT@70@wsj/06/wsj_0655@wsj@en@on: Contra

CHAIN: IDENT@75@wsj/06/wsj_0655@wsj@en@on
LINKS:
 IDENT@0:8:9@IDENT@75@wsj/06/wsj_0655@wsj@en@on: the weekend
 IDENT@2:6:7@IDENT@75@wsj/06/wsj_0655@wsj@en@on: the weekend

CHAIN: IDENT@71@wsj/06/wsj_0655@wsj@en@on
LINKS:
 IDENT@0:26:28@IDENT@71@wsj/06/wsj_0655@wsj@en@on: the Contra rebels
 IDENT@1:13:14@IDENT@71@wsj/06/wsj_0655@wsj@en@on: the Contras
 IDENT@2:15:21@IDENT@71@wsj/06/wsj_0655@wsj@en@on: the rebels seeking
*PRO* to topple him
 IDENT@2:25:26@IDENT@71@wsj/06/wsj_0655@wsj@en@on: the Contras
 IDENT@2:33:33@IDENT@71@wsj/06/wsj_0655@wsj@en@on: they
 IDENT@2:40:40@IDENT@71@wsj/06/wsj_0655@wsj@en@on: their
 IDENT@7:27:28@IDENT@71@wsj/06/wsj_0655@wsj@en@on: the Contras
 IDENT@8:16:17@IDENT@71@wsj/06/wsj_0655@wsj@en@on: the rebels
 IDENT@11:12:13@IDENT@71@wsj/06/wsj_0655@wsj@en@on: the Contras
 IDENT@19:19:20@IDENT@71@wsj/06/wsj_0655@wsj@en@on: the Contras
 IDENT@19:34:35@IDENT@71@wsj/06/wsj_0655@wsj@en@on: the Contras
 IDENT@20:16:17@IDENT@71@wsj/06/wsj_0655@wsj@en@on: the Contras
 IDENT@20:23:23@IDENT@71@wsj/06/wsj_0655@wsj@en@on: themselves
 IDENT@26:11:12@IDENT@71@wsj/06/wsj_0655@wsj@en@on: the Contras
 IDENT@27:6:6@IDENT@71@wsj/06/wsj_0655@wsj@en@on: they


CHAIN: IDENT@69@wsj/06/wsj_0655@wsj@en@on
LINKS:
 IDENT@0:0:3@IDENT@69@wsj/06/wsj_0655@wsj@en@on: Nicaraguan President
Daniel Ortega
 IDENT@0:11:11@IDENT@69@wsj/06/wsj_0655@wsj@en@on: his
 IDENT@2:1:3@IDENT@69@wsj/06/wsj_0655@wsj@en@on: Mr. Ortega 's
 IDENT@2:21:21@IDENT@69@wsj/06/wsj_0655@wsj@en@on: him
 IDENT@3:22:24@IDENT@69@wsj/06/wsj_0655@wsj@en@on: Mr. Ortega 's
 IDENT@6:0:1@IDENT@69@wsj/06/wsj_0655@wsj@en@on: Mr. Ortega
```

```
 IDENT@6:6:6@IDENT@69@wsj/06/wsj_0655@wsj@en@on: he
 IDENT@6:21:21@IDENT@69@wsj/06/wsj_0655@wsj@en@on: his
 IDENT@7:2:2@IDENT@69@wsj/06/wsj_0655@wsj@en@on: he
 IDENT@7:4:4@IDENT@69@wsj/06/wsj_0655@wsj@en@on: his
 IDENT@7:10:10@IDENT@69@wsj/06/wsj_0655@wsj@en@on: he
 IDENT@8:0:0@IDENT@69@wsj/06/wsj_0655@wsj@en@on: He
 IDENT@9:3:5@IDENT@69@wsj/06/wsj_0655@wsj@en@on: Mr. Ortega 's
 IDENT@13:16:18@IDENT@69@wsj/06/wsj_0655@wsj@en@on: Mr. Ortega 's
 IDENT@14:27:27@IDENT@69@wsj/06/wsj_0655@wsj@en@on: his
 IDENT@14:29:31@IDENT@69@wsj/06/wsj_0655@wsj@en@on: Mr. Ortega 's
 IDENT@15:22:24@IDENT@69@wsj/06/wsj_0655@wsj@en@on: the Nicaraguan
leader
 IDENT@16:27:29@IDENT@69@wsj/06/wsj_0655@wsj@en@on: Mr. Ortega 's
 IDENT@20:0:2@IDENT@69@wsj/06/wsj_0655@wsj@en@on: Mr. Ortega 's
 IDENT@24:14:16@IDENT@69@wsj/06/wsj_0655@wsj@en@on: Mr. Ortega 's
 IDENT@24:33:33@IDENT@69@wsj/06/wsj_0655@wsj@en@on: he
 IDENT@25:10:11@IDENT@69@wsj/06/wsj_0655@wsj@en@on: Mr. Ortega
 IDENT@25:18:18@IDENT@69@wsj/06/wsj_0655@wsj@en@on: he
 IDENT@25:29:29@IDENT@69@wsj/06/wsj_0655@wsj@en@on: he
 IDENT@25:38:38@IDENT@69@wsj/06/wsj_0655@wsj@en@on: his
 IDENT@26:4:5@IDENT@69@wsj/06/wsj_0655@wsj@en@on: Mr. Ortega


----------------------------------------------------
-------------------- NAMES -------------------------
----------------------------------------------------

<name_entity_set object:
      <name_entity object: id: GPE@0:12:12@wsj/06/wsj_0655@wsj@en@on
name_entity_type: GPE; sentence_id: 0; start_token_index: 12;
end_token_index: 12; ne_string: 'U.S.';>
      <name_entity object: id: NORP@0:0:0@wsj/06/wsj_0655@wsj@en@on
name_entity_type: NORP; sentence_id: 0; start_token_index: 0;
end_token_index: 0; ne_string: 'Nicaraguan';>
      <name_entity object: id:
ORGANIZATION@0:24:24@wsj/06/wsj_0655@wsj@en@on name_entity_type:
ORGANIZATION; sentence_id: 0; start_token_index: 24; end_token_index:
24; ne_string: 'Contra';>
>
======================================================================
```

For each sentence, the ONF form begins with the sentence and the parse tree. Following the parse tree, each word appears on a line by itself, with its token ID number and its OntoNotes wordsense, if one has been assigned.

For verbs or other predicate words, the line for the word is followed by a block that specifies the predicate and its arguments. Each argument (ARG0, ARG1, ARGM-MOD, etc.) is specified in a “word:height” format that specifies the token number of the first word in the argument and the number of levels up in the tree to go to find the appropriate node. For example, in the “accomplish” predicate for word 6 in the above example, the ARG0 is “0:1”, the NP-SBJ node that is one level up from word 0 in the sentence, which is “Nicaraguan”.

At the end of each file, the coreference chains are specified, using a “sentence:word:height” format. In the above example, the chains that include an element

from the example sentence are shown, which link “Ortega”, “Contra”, “the Contra rebels”, and “the weekend” to later mentions in subsequent sentences in the document.

The ONF for each file also includes a sentence by sentence listing of the entity names in the document.

6.4 The Treebank View

The Treebank view uses the same parenthesized format at the original Penn Treebank2.

```
((S (S-ADV (NP-SBJ (-NONE- *PRO*))
    (VP (VBG Judging)
        (PP-CLR (IN from)
           (NP (NP (DT the) (NNS Americana))
               (PP-LOC (IN in)
                   (NP (NP (NNP Haruki) (NNP Murakami) (POS 's))
                       (`` ``)
                       (NX-TTL (NP (DT A) (NNP Wild) (NNP Sheep) (NNP Chase)))
                        ('' '')
                        (NP (-LRB- -LRB-)
                            (NP (NNP Kodansha))
                                (, ,)
                                (NP (CD 320) (NNS pages))
                                (, ,)
                                (NP ($ $)
                                    (CD 18.95)
                                    (-NONE- *U*))
                       (-RRB- -RRB-))))))))
     (, ,)
     (NP-SBJ (NP (NN baby) (NNS boomers))
        (PP-LOC (IN on)
           (NP (NP (DT both) (NNS sides))
               (PP (IN of)
                   (NP (DT the) (NNP Pacific))))))
     (VP (VBP have)
         (NP (NP (DT a) (NN lot))
             (PP (IN in)
                 (NP (NN common)))))
     (. .)))
```

6.5 Proposition Bank View

In the PropBank view, each line of data contains information about the predicate argument structures of a particular verb instance. The elements are represented using space-separated columns, as follows:

wsj-filename sentence terminal tagger frameset inflection proplabel proplabel

The content of each column is described in detail below, with both English and Chinese examples given.

- wsj-filename: the name of the file in merged Penn Treebank, WSJ section, or in the Penn Chinese Treebank.
- sentence: the number of the sentence in the file (starting with 0)
- terminal: the number of the terminal in the sentence that is the location of the verb. Note that the terminal number counts empty constituents as terminals and starts with 0. This will hold for all references to terminal number in this

description.
In the English example:
(NP-1 (NN John) (VP (VB wants) (S (NP (-NONE- *-1)) (VP (TO to) (V swim)))))
the terminal numbers are: John 0; wants 1; *-1 2; to 3; swim 4
In the Chinese example:
(IP (NP-SBJ (DNP (NP (NN 货币)(NN 回笼))(DEG 的))(NP (NN 增加)))(PU ，)
(VP (PP-BNF (P 为)(IP (NP-SBJ (-NONE- *PRO*))(VP (VV 平抑)(NP-OBJ (NP (DP (DT 全))
(NP (NN 区)))(NP (NN 物价))))))(VP (VV 发挥)(AS 了)(NP-OBJ (NN 作用))))(PU 。))
the terminal numbers are:
货币 0 回笼 1 的 2 增加 3 ， 4 为 5 *PRO* 6 平抑 7 全 8 区 9 物价 10 发挥 11 了 12 作用 13 。 14

- tagger: the name of the annotator, or "gold" if it's been double annotated and adjudicated.
- Frameset: The frameset identifier from the frames file of the verb. For example, 'dial.01' refers to the frames file for 'dial', (frames/dial.xml) and the roleset element in that frames file whose attribute 'id' is 'dial.01'.
  There are some instances which have yet to be disambiguated, these are marked as 'lemma.XX'.
  For Chinese, the names of the frame files are composed of numerical id, plus the pinyin of the verb. The numerical ids can be found in the enclosed verb list (verbs.txt).
- Inflection: Used only in the English data, the inflection field consists of 5 characters representing person, tense, aspect, voice, and form of the verb, respectively.
  Each of the characters may be '-', representing 'none'.
  The possible values of each of the fields character codes are as follows:
  - form: i=infinitive g=gerund p=participle v=finite
  - tense: f=future p=past n=present
  - aspect: p=perfect o=progressive b=both perfect and progressive
  - person: 3=3rd person
  - voice: a=active p=passive
- proplabel (a.k.a. “arglabel”): A string representing the annotation associated with a particular argument or adjunct of the proposition. Each proplabel is dash '-' delimited and has columns for (1) the syntactic relation, (2) the label, and (3) optional argument features. The contents of these columns are described in detail in the following paragraphs.

Element (1) of the proplabel for each proposition specifies the syntactic relation. This can be in one of 4 forms:

- form 1: <terminal number>:<height>
  A single node in the syntax tree of the sentence in question, identified by the first terminal the node spans together with the height from that terminal to the syntax node (a height of 0 represents a terminal).
  For example, in the sentence
  (S (NP-1 (NN John) (VP (VB wants) (S (NP (-NONE- *-1)) (VP (TO to) (V swim)))))
  A syntactic relation of "2:1" represents the NP immediately dominating the terminal "(-NONE- *-1)" and a syntactic relation of "0:2" represents the "S" node.
  In the Chinese sentence
  (IP (NP-TPC (DP (DT 这些))(CP (WHNP-1 (-NONE- *OP*)) (CP (IP (NP-SBJ (-NONE- *T*-1)) (VP (ADVP (AD 已))(VP (VV 开业))))(DEC 的)))(NP (NN 外商)(NN 投资)(NN 企业))) (NP-ADV (NN 绝大部分))(NP-SBJ (NN 生产)(NN 经营)(NN 状况))(VP (ADVP (AD 较)) (VP (VA 好)))(PU 。))
  the address of "1:3" represents the top IP node and 2:2 represents the CP node

- form 2: terminal number:height*terminal number:height*
  A trace chain identifying coreference within sentence boundaries.
  For example in the sentence
  ((NP-1 (NN John) (VP (VB wants) (S (NP (-NONE- *-1)) (VP (TO to) (V swim)))))
  A syntactic relation of "2:1*0:1" represents the NP immediately dominating (-NONE- *-1) and the NP immediately dominating "(NN John)".
  In the Chinese sentence
  (IP (NP-TPC (DP (DT 这些))(CP (WHNP-1 (-NONE- *OP*)) (CP (IP (NP-SBJ (-NONE- *T*-1)) (VP (ADVP (AD 已))(VP (VV 开业))))(DEC 的)))(NP (NN 外商)(NN 投资)(NN 企业))) (NP-ADV (NN 绝大部分))(NP-SBJ (NN 生产)(NN 经营)(NN 状况))(VP (ADVP (AD 较)) (VP (VA 好)))(PU 。))
  the address of of "2:0*1:0*6:1" represents the fact nodes '2:0' (-NONE- *T*-1), '1:0' (-NONE- *OP*) and '6:1' (NP (NN 外商)(NN 投资)(NN 企业)) are coreferential.

- form 3: terminal number:height, terminal number:height,
  A split argument, where there is no single node that captures the argument and the components are not coreferential, eg the utterance in "I'm going to", spoke John, "take it with me". This form is also used to denote phrasal variants of verbs. For example, in the phrase fragment
  (S (NP (NN John)) (VP (VB keeps) (PRT on) (NP ...))
  The phrasal verb "keep_on" would be identified with the syntactic relation "1:0,2:0".

- form 4: terminal number:height,terminal number:height*terminal number:height...
  This form is a combination of forms 2 and 3. When this occurs, the ',' operator is understood to have precedence over the '*' operator. For example, in the sentence

```
(NP (DT a) (NN series) )
   (PP (IN of)(NP (NNS intrigues) ))
     (SBAR
      (WHNP-4 (WDT that) )
      (S
       (NP-SBJ (-NONE- *T*-4) )
       (VP (VBZ has)
        (S
         (NP-SBJ (NN everyone) )
         (VP (VBG fearing)
```

The proplabel 28:1,30:1*32:1*33:0-ARG0 is to be understood as a trace-chain (form 2), one of whose constituents is a split argument (form 3) - i.e. grouped like so: ((28:1,30:1)*32:1*33:0). The interpretation of this argument is that the "causer of action" (ARG0 of have.04) is signified by the following trace-chain: *T*-4 --> that --> ([a series][of intrigues])

Element (2) of the proplabel for each proposition specifies the 'label'.The argument label one of {rel, ARGA, ARGM} + { ARG0, ARG1, ARG2, ... }. The argument labels correspond to the argument labels in the frames files (see ./frames). ARGA is used for causative agents, ARGM for adjuncts of various sorts, and 'rel' refers to the surface string of the verb.

Element (3) of the proplabel for each proposition supplies argument features (optional for numbered arguments; required for ARGM). Argument features can either be a labeled feature, or a preposition. For the English data, the labeled features include:

- EXT - extent
- DIR - direction
- LOC - location
- TMP - temporal
- REC - reciprocal
- PRD - predication
- NEG - negation
- MOD - modal
- ADV - adverbial
- MNR - manner
- CAU - cause
- PNC - purpose not cause.
- DIS - discourse

Preposition features are attached to argument labels when the argument is tagged on a PP node.

For the Chinese data, the following functional tags are used for “split” numbered arguments:

- PSR - possessor
- PSE - possessee
- CRD - coordinator

- PRD - predicate
- QTY - quantity

The propositional tags for numbered arguments are: AT, AS, INTO, TOWARDS, TO, ONTO

The functional tags in the Chinese data for ARGMs are as follows:

- ADV - adverbial, default tag
- BNF - beneficiary
- CND - conditional
- DIR - directional
- DIS - discourse
- DGR - degree
- EXT - extent
- FRQ - frequency
- LOC - location
- MNR - manner
- NEG - negation**
- PRP - purpose and reason
- TMP - temporal
- TPC - topic

Here are some sample lines of OntoNotes output in the PropBank View:

```
wsj/00/wsj_0020.mrg@wsj@en@on 0 4 gold claim.01 ----- 3:0*0:1-ARG0 4:0-
rel 5:2-ARG1

wsj/00/wsj_0020.mrg@wsj@en@on 0 12 gold remove.01 ----- 0:1-ARG0 3:2-
ARGM-ADV 12:0-rel 13:2-ARG1 20:1-ARG2-from

wsj/00/wsj_0020.mrg@wsj@en@on 0 29 gold watch.01 ----- 24:1*25:0-LINK-
SLC 26:1-ARG0 29:0-rel 30:0*25:1-ARG1 31:1-ARGM-CAU

wsj/00/wsj_0020.mrg@wsj@en@on 0 34 gold fail.01 ----- 32:1-ARG1 33:1-
ARGM-ADV 34:0-rel 35:2-ARG2

wsj/00/wsj_0020.mrg@wsj@en@on 0 37 gold honor.01 ----- 35:0*32:1-ARG0
37:0-rel 38:1-ARG1

wsj/00/wsj_0020.mrg@wsj@en@on 1 17 gold remain.01 ----- 0:1-ARGM-DIS
2:2-ARG1 16:0-ARGM-MOD 17:0-rel 18:1-ARG3 24:1-ARGM-CAU

wsj/00/wsj_0020.mrg@wsj@en@on 1 37 gold announce.01 ----- 32:1-ARG0
37:0-rel 38:1-ARG1

wsj/00/wsj_0020.mrg@wsj@en@on 2 10 gold face.01 ----- 0:1-ARGM-LOC 7:1-
ARG0 9:0-ARGM-MOD 10:0-rel 11:2-ARG1 18:1-ARGM-ADV
```

6.6 Word Sense View

The word sense annotation view is formatted with one line per annotated word instance. That line specifies the file, sentence number, word number, lemma, and the selected sense, as defined in the sense inventory file for that lemma. (The "?" placeholders fill slots in the format that were used internally by the word sense annotation tool.)

Here are some sample lines of output in the word sense view:

```
wsj/02/wsj_0242@wsj@en@on 0 4 complain-v ?,? 1
wsj/02/wsj_0242@wsj@en@on 0 9 push-v ?,? 1
wsj/02/wsj_0242@wsj@en@on 0 15 create-v ?,? 1
wsj/02/wsj_0242@wsj@en@on 0 25 affect-v ?,? 1
wsj/02/wsj_0242@wsj@en@on 1 3 pace-n ?,? 1
wsj/02/wsj_0242@wsj@en@on 1 11 aim-v ?,? 2
wsj/02/wsj_0242@wsj@en@on 1 24 reduction-n ?,? 1
wsj/02/wsj_0242@wsj@en@on 1 36 register-v ?,? 1
wsj/02/wsj_0242@wsj@en@on 2 3 call-v ?,? 3
wsj/02/wsj_0242@wsj@en@on 2 6 agreement-n ?,? 1
```

6.7 Coreference View

The coreference view is formatted using in-line annotation. COREF tags are used to mark the beginning and end of constituents that should be linked, with ID number attributes defining the chains. The TYPE attribute distinguishes the normal IDENT coref from the special APPOS type used for appostitives.

The text that underlies the coreference view follows the Treebank tokenization, and also includes the trace and empty category elements (like “*”, “*-2”, and “*U*”) found in the Treebank analysis, since those can also participate in the coreference chains.

```
<DOC>
<DOCNO>wsj_0037.mrg</DOCNO>
* Judging from the Americana in <COREF ID="135" TYPE="IDENT"><COREF
ID="144" TYPE="IDENT">Haruki Murakami 's</COREF> `` A Wild Sheep Chase
'' -LRB- <COREF ID="140" TYPE="IDENT">Kodansha</COREF> , 320 pages , $
18.95 *U* -RRB-</COREF> , baby boomers on both sides of the Pacific
have a lot in common .
Although *-2 set *-1 in <COREF ID="137" TYPE="IDENT">Japan</COREF> ,
<COREF ID="135" TYPE="IDENT">the novel 's</COREF> texture is almost
entirely Western , especially American .
<COREF ID="17" TYPE="IDENT">Characters</COREF> drink Salty Dogs ,
whistle `` Johnny B. Goode '' and watch Bugs Bunny reruns .
<COREF ID="17" TYPE="IDENT">They</COREF> read Mickey Spillane and talk
about Groucho and Harpo .
<COREF ID="17" TYPE="IDENT">They</COREF> worry about <COREF ID="17"
TYPE="IDENT">their</COREF> careers , drink too much and suffer through
broken marriages and desultory affairs .
This is <COREF ID="137" TYPE="IDENT">Japan</COREF> ?
...
After years of decline , <COREF ID="22" TYPE="IDENT">weddings in
France</COREF> showed a 2.2 % upturn <COREF ID="170" TYPE="IDENT">last
year</COREF> , with 6,000 more couples *ICH*-1 <COREF ID="22"
TYPE="IDENT">exchanging</COREF> rings in <COREF ID="170"
```

```
TYPE="IDENT">1988</COREF> than in <COREF ID="171" TYPE="IDENT">the
previous year</COREF> , the national statistics office said 0 *T*-2 .

But <COREF ID="180" TYPE="APPOS" SUBTYPE="ATTRIBUTE">the number</COREF>
of <COREF ID="22" TYPE="IDENT">weddings</COREF> <COREF ID="170"
TYPE="IDENT">last year</COREF> -- <COREF ID="180" TYPE="APPOS"
SUBTYPE="HEAD">271,124</COREF> -- was still well below the 400,000
registered * in <COREF ID="172" TYPE="APPOS"-
SUBTYPE="HEAD">1972</COREF> , <COREF ID="172" TYPE="APPOS"
SUBTYPE="ATTRIBUTE">the last year of <COREF ID="22"
TYPE="IDENT">increasing marriages</COREF></COREF> .

</DOC>
```

6.8 Entity Names View

The entity names annotation view is formatted using in-line ENAMEX markup.

Here is a portion of a sample document in the entity names view:

```
<DOC>

<DOCNO>wsj/02/wsj_0242@wsj@en@on</DOCNO>

Some <ENAMEX TYPE="GPE">U.S.</ENAMEX> allies are complaining that
President <ENAMEX TYPE="PERSON">Bush</ENAMEX> is pushing conventional-
arms talks too quickly , creating a risk that negotiators will make
errors that could affect the security of Western Europe for <ENAMEX
TYPE="DATE">years</ENAMEX> .

Concerns about the pace of the <ENAMEX TYPE="GPE">Vienna</ENAMEX> talks
-- which are aimed at the destruction of some 100,000 weapons , as well
as major reductions and realignments of troops in central <ENAMEX
TYPE="LOCATION">Europe</ENAMEX> -- also are being registered at the
<ENAMEX TYPE="ORGANIZATION">Pentagon</ENAMEX> .

Mr. <ENAMEX TYPE="PERSON">Bush</ENAMEX> has called for an agreement by
next September at the latest .

But some <ENAMEX TYPE="NORP">American</ENAMEX> defense officials
believe the North Atlantic Treaty Organization should take more time to
examine the long-term implications of the options being considered .

For <ENAMEX TYPE="CARDINAL">one</ENAMEX> thing , <ENAMEX
TYPE="ORGANIZATION">Pentagon</ENAMEX> officials , who asked not to be
identified , worry that the <ENAMEX TYPE="GPE">U.S.</ENAMEX> will have
a much tougher time persuading <ENAMEX TYPE="NORP">Europeans</ENAMEX>
to keep some short-range nuclear weapons on their soil once <ENAMEX
TYPE="NORP">Soviet</ENAMEX> armored forces are thinned out .

...

</DOC>
```

6.9 Ontology View

During OntoNotes annotation, the information that connects the word senses with the ontology is stored as a number of separate XML files, which are the source from which the ontology information gets loaded into the database. The ontology upper model is stored as the single large XML file “upper-model.xml”, which represents the toplevel

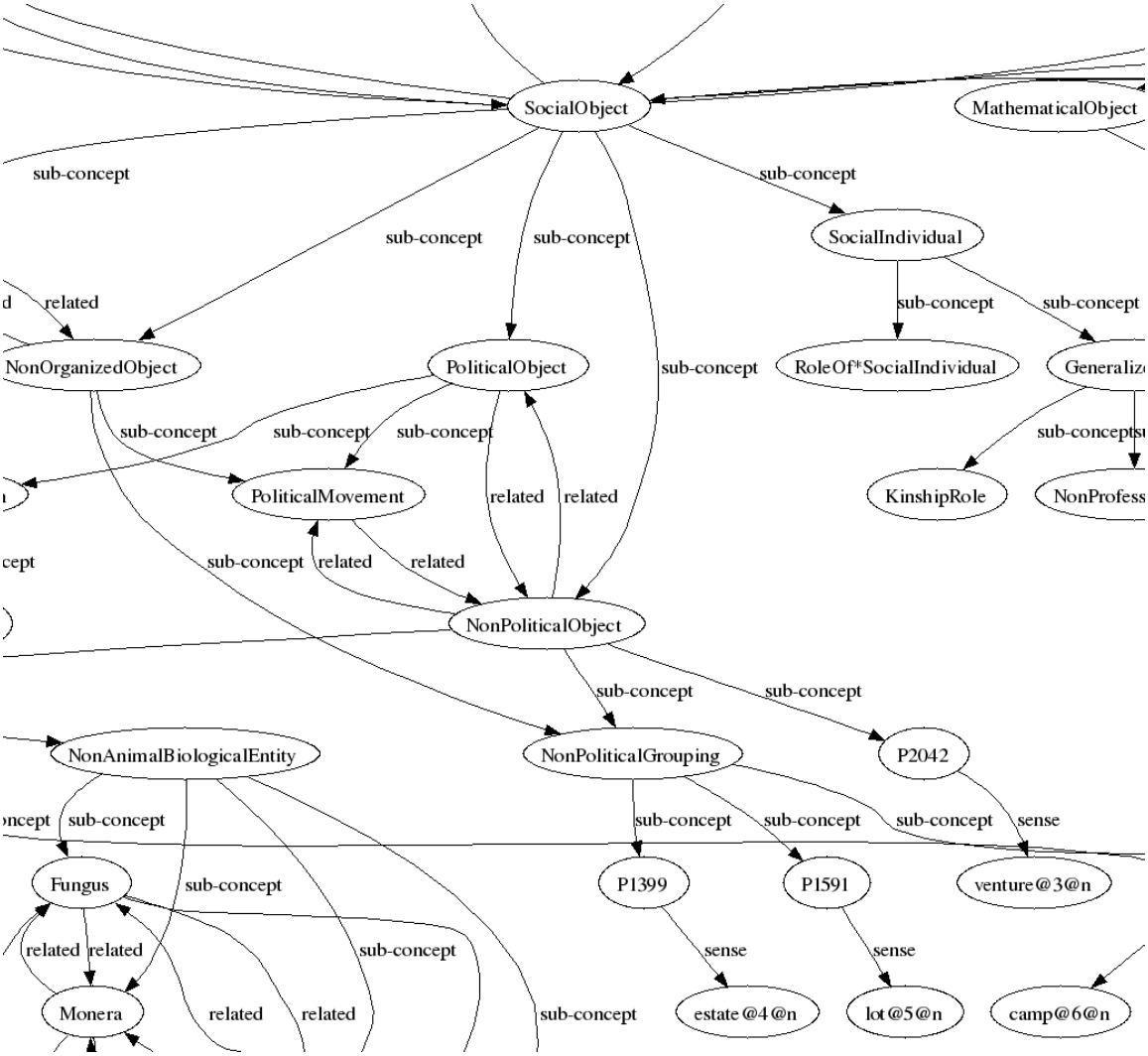
concepts with their interconnections. The sense pools created from the word sense annotation are represented, one-per-file, in XML files in the "sense-pools" sub-directory.

To enable easier visualization and interpretation, the ontology view that can be generated from the OntoNotes database comes in the form of source files for an open source graphics package (Graphviz) which can then display the ontology as an actual tree structure. that represented as a .dot file. The Graphviz package, available at www.graphviz.org (we used version 2.14), uses a ".dot" file format to encode the nodes and arcs of the graph. A portion of the .dot file for the OntoNotes ontology is shown below, where lines containing "->" encode arcs, and the other encode nodes:

```
digraph O {
    "Object" [id="Object", commentary="Entities that are 'constant' ...];
    "Belief" [id="Belief", commentary=""];
    "Artifact" [id="Artifact", commentary="physical objects intentionally ...];
    "BiologicalObject" [id="BiologicalObject", commentary="Objects that have life"];
    "Animal" [id="Animal", commentary="the kingdom Animalia: volitional living things ...
];
    ...
    ...
    "Object" -> "IntangibleObject" [label="sub-concept"];
    "Object" -> "Quality" [label="related"];
    "Object" -> "TangibleObject" [label="sub-concept"];
    "Object" -> "UnrootedObject" [label="sub-concept"];
    "Belief" -> "Intention" [label="related"];
    "Belief" -> "Perception" [label="related"];
    "Belief" -> "Thought" [label="related"];
    "Artifact" -> "NaturalNonLivingObject" [label="related"];
    "BiologicalObject" -> "Animal" [label="sub-concept"];
    "BiologicalObject" -> "NonBiologicalObject" [label="related"];
    "BiologicalObject" -> "NonVolitionalBiologicalObject" [label="sub-concept"];
    "Animal" -> "Invertebrate" [label="sub-concept"];
    "Animal" -> "NonVolitionalBiologicalObject" [label="related"];
    "Animal" -> "RoleOf\*Animal" [label="sub-concept"];
    "Animal" -> "Vertebrate" [label="sub-concept"];
    ...
    ...
    "Animal" -> "P2617" [label="pool"];
    "Artifact" -> "P1081" [label="pool"];
    "Artifact" -> "P1083" [label="pool"];
    "Artifact" -> "P1084" [label="pool"];
    "Artifact" -> "P1085" [label="pool"];
    ...
    ...
    "P0071" -> "endeavor@2@n" [label="sense"];
```

```
    "P0071" -> "experiment@2@n" [label="sense"];
    "P0071" -> "gamble@1@n" [label="sense"];
    "P0071" -> "undertaking@1@n" [label="sense"];
    "P0071" -> "venture@1@n" [label="sense"];
    "P1044" -> "arsenal@1@n" [label="sense"];
    "P1045" -> "arsenal@2@n" [label="sense"];
    "P1045" -> "magazine@4@n" [label="sense"];
    "P1046" -> "arsenal@3@n" [label="sense"];
    "P1076" -> "bar@1@n" [label="sense"];
    "P1077" -> "bar@10@n" [label="sense"];
    ...
    ...
}
```

Graphviz provides several ways of visualizing the graph encoded by a .dot file. One option is to generate an image file in any of the common image file formats. The following figure shows a portion of the ontology graph when generated in .png format. The nodes with numeric suffixes represent word senses or sense pools; nodes without such suffixes represent concept nodes from the ontology's upper model. Not only does this simplify the vizualization of the ontology, but there are standard libraries, for example, Boost C++ library (www.boost.org), that can read the .dot file and generate the data structure for easy manipulation and integration into an application.



6.10 Supplementary Data

The interpretation of certain values in the annotation database is specified in supplementary data files included in the data directory of the distribution. The PropBank frames files specify the pattern and meaning of the propositional argument labels, and the word sense inventory files specify the set of possible meanings for each word.

6.10.1 PropBank Frame Files

The argument numbering used in the PropBank annotation is interpreted in the frames files. The frames file for each verb specifies one or more frames, and each frame defines a particular set of arguments with their interpretation. The data directory of this distribution includes separate Chinese and English directories containing the frames files for each of the verbs covered.

6.10.2 Sense Inventory Files

The sense inventory files specify the range of possible word senses for each annotated noun and verb. Each word sense is described with examples, and the meanings are also characterized in terms of a set of primitive semantic features like "+concrete", "+animate", etc.

The inventory files are XML documents, with the entry for each lemma organized as a sequence of senses. Each sense has a number, a name attribute that provides a short definition, a list of examples, and a set of mappings that relate the sense back to a WordNet or a ProbBank frame, as appropriate.

The sense inventory files are included in the data directory, organized by language and by part of speech.

6.11 Access Script Documentation

The database contains the merged layers of annotation for both the English and Chinese corpora. It was created by loading the separate Treebank, PropBank, word sense, and coreference sources and merging them into a set of linked relational database tables. A dump-file image of the resulting database is included in this distribution, along with the original source files and the code that was used to do the merge.

Code is also provided to extract views from the merged database. In particular, each of the original source-file formats is defined as a view that can be extracted from the database. (In a couple cases, there are minor formatting differences between the original source files and the view file; in those cases, both versions are included.) Another predefined view is the "OntoNotes Normal Form" view, a textual version of the combined annotation, intended for human review. As an alternative, the OntoViewer utility, included with this release, can be used to provide a flexible interactive view of the various annotation layers, including a propositions view that shows the nested structure of the multiple propositions in a sentence.

Users can also define their own SQL queries to search for particular constructions or contexts in the combined corpus annotations, or can use the data access primitives provided to define their own views of the data.

Documentation describing the database schema and API, the loading routines, and the access scripts can be found in "OntoNotes DB Tool" guide in the documentation directory.

7 References

A. Abeille (ed.). 2003. *Treebanks: Building and Using Parsed Corpora.* Kluwer Academic Publishers.

H. Alshawi. 1992. *The CORE Language Engine*. MIT Press.

O. Babko-Malaya, A. Bies, A. Taylor, S. Yi, M. Palmer, M. Marcus, S. Kulick and L. Shen. Issues in Synchronizing the English Treebank and Propbank. *Proceedings of the Workshop on Frontiers in Linguistically Annotated Corpora*.

S. Bangalore and A. Joshi. 1999. Supertagging: An Approach to Almost Parsing. *Computational Linguistics 25*, pp. 237-265.

R. Belvin. 1993. The two causative *haves* are the two possessive *haves*. *Proceedings of the Chicago Linguistics Society (CLS-29).*

E. Charniak, 2000. A Maximum-Entropy-Inspired Parser. *Proceedings of the North American Association for Computational Linguistics (NAACL-2000).*

J. Chen, A. Schein, L. Ungar and M. Palmer. 2006. An Empirical Study of the Behavior of Word Sense Disambiguation. *Proceedings of NAACL-HLT 2006*.

M. Collins, 1998. Three Generative Lexicalized Models for Statistical Parsing. *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics (ACL).*

M. Collins, 2000. Discriminate Reranking for Natural Language Parsing. *International Conference on Machine Learning (ICML).*

C. Fellbaum (ed.) 1998. *WordNet: An On-line Lexical Database and Some of its Applications*. MIT Press, Cambridge, MA.

J. Hobbs et al. 1997. FASTUS: A Cascaded Finite-State Transducer for Extraction Information from Natural-Language Text. In Roche and Schabes, ed., *Finite State Devices for Natural Language Processing*, MIT Press.

E. H. Hovy, M. Marcus, M. Palmer, S. Pradhan, L. Ramshaw, and R. Weischedel. 2006. OntoNotes: The 90% Solution. *Proceedings of the Human Language Technology / North American Association of Computational Linguistics conference (HLT-NAACL 2006).* New York, NY.

K. Knight and S. K. Luk. 1994. Building a Large-Scale Knowledge Base for Machine Translation. *Proceedings of AAAI-94* (Seattle, WA, 1994)

B. Levin. 1993. *English Verb Classes and Alternations: A Preliminary Investigation*. University of Chicago Press, Chicago, IL.

K. Mahesh. 1996. Ontology Development for Machine Translation: Ideology and Methodology. New Mexico State University CRL report MCCS-96-292.

M. Marcus, M. Marcinkiewicz, and B. Santorini. 1993. Building a Large Annotate Corpus of English: The Penn Treebank. *Computational Linguistics 19.*

G. Miller. 1990. WordNet: An online lexical database. *International Journal of Lexicography*, 3(4).

S. Miller, L. Ramshaw, H. Fox, and R. Weischedel. 2000. A Novel Use of Statistical Parsing to Extract Information from Text. *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL).*

T. O'Hara, K. Mahesh, and S. Nirenburg. 1998. Lexical acquisition with WordNet and the Mikrokosmos Ontology. *Proceedings of the COLING/ACL Workshop on Usage of WordNet in Natural Language Processing Systems*. Montreal, Canada.

M. Palmer, O. Babko-Malaya, and H. T. Dang. 2004. Different Sense Granularities for Different Applications. *Second Workshop on Scalable Natural Language Understanding Systems, at HLT/NAACL-04.*

M. Palmer, H. Dang, and C. Fellbaum. 2006. Making Fine-grained and Coarse-grained Sense Distinctions, Both Manually and Automatically. *Journal of Natural Language Englineering.*

A. Philpot, E.H. Hovy, and P. Pantel. 2005. The Omega Ontology. *Proceedings of the ONTOLEX Workshop at the International Conference on Natural Language Processing (IJCNLP).* Jeju Island, Korea.

A. Ratnaparkhi. 1997. A Linear Observed Time Statistical Parser Based on Maximum Entropy Models. *Second Conference on Empirical Methods in Natural Language Processing (EMNLP).* Providence, Rhode Island.

N. Xue, F. Xia, F-D. Chiou and M. Palmer. 2005. The Penn Chinese TreeBank: Phrase Structure Annotation of a Large Corpus. *Natural Language Engineering, 11(2)207-238.*

N. Xue and M. Palmer. 2007. Adding Semantic Roles to the Chinese Treebank. Under review at *Natural Language Engineering*.

Yu, L.C., C.H. Wu, A. Philpot, and E.H. Hovy. 2007. OntoNotes: Sense Pool Verification Using Google N-gram and Statistical Tests. *Proceedings of the OntoLex Workshop at the 6th International Semantic Web Conference (ISWC 2007)*. Busan, Korea.