OntoNotes Release 5.0

with OntoNotes DB Tool v0.999 beta

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

2012-09-28

Ralph Weischedel, Sameer Pradhan, Lance Ramshaw, Jeff Kaufman, Michelle Franchini, Mohammed El-Bachouti

Nianwen Xue

Martha Palmer, Jena D. Hwang, Claire Bonial, Jinho Choi, Aous Mansouri, Maha Foster and Abdel-aati Hawwary

Mitchell Marcus, Ann Taylor, Craig Greenberg

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

Contents

1 Introduction

This document describes the final release (v5.0) of OntoNotes, an annotated corpus whose development was supported under the GALE program of the Defense Advanced Research Projects Agency, Contract No. HR0011-06-C-0022. The annotation is provided both in separate text files for each annotation layer (Treebank, PropBank, word sense, etc.) and in the form of an integrated relational database with a Python API to provide convenient cross-layer access. 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 database are included in the documentation directory of the distribution.

1.1 Summary Description of the OntoNotes Project

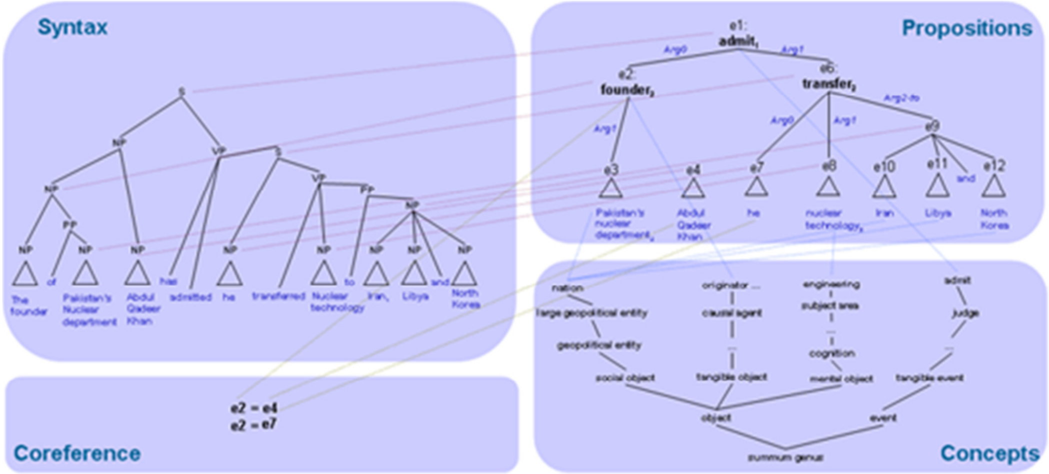
Natural language applications like machine translation, question answering, and summarization currently are forced to depend on impoverished text models like bags of words or n-grams, while the decisions that they are making ought to be based on the meanings of those words in context. That lack of semantics 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 was a collaborative effort between BBN Technologies, Brandeis University, the University of Colorado, the University of Pennsylvania, and the University of Southern California's Information Sciences. The goal was to annotate a large corpus comprising various genres (news, broadcast, talk shows, weblogs, usenet newsgroups, and conversational telephone speech) 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 adds coreference to PropBank, and includes partial word sense disambiguation for some nouns and verbs, with the word senses connected to an ontology. OntoNotes includes roughly 1.5 million words of English, 800 K of Chinese, and 300 K of Arabic. More details are provided in Weischedel et al. (2011)

This resource is being made 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 strove to ensure that each layer of annotation in OntoNotes have at least 90% inter-annotator agreement..

This level of semantic representation goes far beyond the entity and relation types targeted in the ACE program, since every concept in the text is 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, since it is the first semantic resource of this substantial size ever produced. As demonstrated with the Treebank and WordNet, a publicly available resource can unleash an enormous amount of work internationally on algorithms and on the automated creation of semantic resources in numerous other domains and genres. We hope 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.

1.2 Corpus and GALE Project Plans

The goal for OntoNotes was to achieve substantial coverage in various genres and in all three GALE languages. The current 5.0 release covers newswire, broadcast news, broadcast conversation, and web data in English and Chinese, a pivot corpus in English, and newswire data in Arabic[1].

[1] For simplicity, the numbers in this table are rounded to the nearest 1000

	English	Chinese	Arabic
Newswire	625 K	250 K	300 K
Broadcast News	200 K	250 K	–
Broadcast Conversation	200 K	150 K	–
Web text	300 K	150 K	–
Telephone conversation	120 K	100 K	–
Pivot Corpus (NT, OT)	300 K	-	–

The English newswire portion includes 300K of English Wall St. Journal newswire and 325K of the English side of the English-Chinese Parallel Treebank (ECTB)[illegible].

There is 200K of English broadcast news data (a part of the TDT-4 collection), and 200K of English broadcast conversation data, half of which is Chinese parallel data, with 50K being English translated from Chinese and 50K Chinese translated from English. There is 200K English web data, 55K of which is translated from Arabic and 75K translated from Chinese. There is 145K English P2.5 data, of which 80K is translated from Chinese and the rest is translated from Arabic. There is approximately 35K P2.5 data for each of Newswire, Broadcast News, Broadcast Conversation, and Web. There is also 85K of English web data consisting of single sentences selected to improve sense coverage. Ontonotes v. 5.0 also included two new English subcorpora: i) A pivot corpus comprising 250K English translation of the New Testament annotated with parse, proposition, name and coreference; and about 100K parses for a portion of the Old Testament; and ii) A telephone conversation corpus (CallHome) comprising about 100K words and annotated with parse, propositions and coreference.

For Chinese, the newswire portion includes 254K of the Chinese side of the English-Chinese Parallel Treebank (ECTB), broadcast news includes 269K of TDT-4 Chinese data, and broadcast conversation includes 169K of data from the LDC's GALE collection. There is also 110K Web data, 40K P2.5 data, and 55K Dev09. Along with the ECTB data and the P2.5 data, 105K of the broadcast conversation data is also parallel, with 50K of the originally-Chinese data having been translated into English and another 55K of the Chinese data have been translated from originally-English data. Further, the 110K of Web data consists of 40K parallel Chinese origin data and 70K parallel English origin data. OntoNotes 5.0 also includes a 100K corpus of telephone conversations (CallHome) annotated with parse, proposition and coreference.

For Arabic, the current release includes 300K of Arabic An-Nahar newswire, with Treebank, word sense, proposition, coreference, and named entity annotation layers. The word sense layer covers only the first 200K, coreference and named entity annotation the first 300K, and Treebank and proposition annotation the full 400K. In OntoNotes 5.0, coverage was extended to verbs with fewer examples in the OntoNotes corpus, and framesets were also created for some verbs that have no instances in the OntoNotes corpora.

[illegible]

Annotation priorities during the final two years of OntoNotes focused particularly on extending PropBank coverage in all three languages and word sense coverage particularly for verbs in English, as well as on providing parallel Treebank and PropBank data for training Machine Translation and Distillation models.

The following tables show the verb proposition coverage and the word sense coverage for nouns and verbs and in all three languages. A couple things to note:

i) In Chinese, words with part of speech other than noun or verb in the Treebank have also been sense tagged since they behave like verbs or nouns in the data. We have not counted those in this table.

ii) In computing word sense coverage we include monosemous words. To determine whether a word is monosemous, we check the number of senses in the OntoNotes inventory files. If the word does not have a sense inventory file, then, for English we use the sense information from WordNet 3.0. (A word that is monosemous in WordNet 3.0 is quite likely to be so in OntoNotes as well, since OntoNotes word senses are generally derived by merging WordNet senses instead of splitting them.) Note that the cells where no noun annotation was performed have been left blank, although for consistency's sake, they could have been filled in based on the number of monosemous nouns at the initial stage of annotation.

	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]
[illegible]			
[illegible]	[illegible]	[illegible]	[illegible]
	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]
	[illegible]	[illegible]	[illegible]
	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]
	[illegible]	[illegible]	[illegible]
	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]
	[illegible]	[illegible]	[illegible]
	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]
	[illegible]	[illegible]	[illegible]
	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]
	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]
	[illegible]	[illegible]	[illegible]
	[illegible]	[illegible]	[illegible]

[illegible]

	[illegible]	[illegible]	
	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]
	[illegible]	[illegible]	[illegible]
	[illegible]	[illegible]	[illegible]

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 provide crucial links to the proposition and coreference annotation in the OntoNotes analysis.

The English Treebank (http://www.cis.upenn.edu/~treebank/) is being developed at the University of Pennsylvania under the supervision of Prof. Mitchell Marcus. Within the OntoNotes project, the University of Pennsylvania does Treebank annotation for the new genres of English text, while also contributing towards improving statistical parsing technology.

In the months leading up to this OntoNotes 5.0 release, work has been underway to make the different English Treebank corpora that are available from OntoNotes and from the LDC more consistent. As part of that effort, the LDC has revised their trees for the English side of the English Chinese Treebank (ECTB), which OntoNotes (gratefully) uses as the foundation for our later layers of annotation for those corpora. The changes included ones related to the “Treebank/PropBank merge”, a set of modifications that were made a while back in both the Treebank and PropBank guidelines to make the two levels of annotation more consistent, as well as consistency changes recommended by the GALE program’s Banks Advisory Committee.

The Chinese Treebank (http://verbs.colorado.edu/chinese/ctb.html) is being developed under the supervision of Prof. Martha Palmer at the University of Colorado and Nianwen Xue at Brandeis University, who also provide the Chinese Treebank annotation for OntoNotes.

The Arabic Treebank is being developed by the Linguistic Data Consortium (http://www.ldc.upenn.edu/). While additional Arabic Treebanking is not part of OntoNotes, our OntoNotes annotation layers for Arabic depend on the parses supplied by the Linguistic Data Consortium at LDC (http://ldc.upenn.edu) under the supervision of Mohamed Maamouri.

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, consider 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:

- ...[[illegible] the company] to ... offer [[illegible] a 15% to 20% stake] [[illegible] to the public]. (wsj 0345)
- ... [[illegible] Sotheby's] ... offered [[illegible] the Dorrance heirs] [[illegible] a money-back guarantee] (wsj 1928)
- ... [[illegible] an amendment] offered [[illegible] by Rep. Peter DeFazio] ... (wsj 0107)
- ... [[illegible] Subcontractors] will be offered [[illegible] 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.

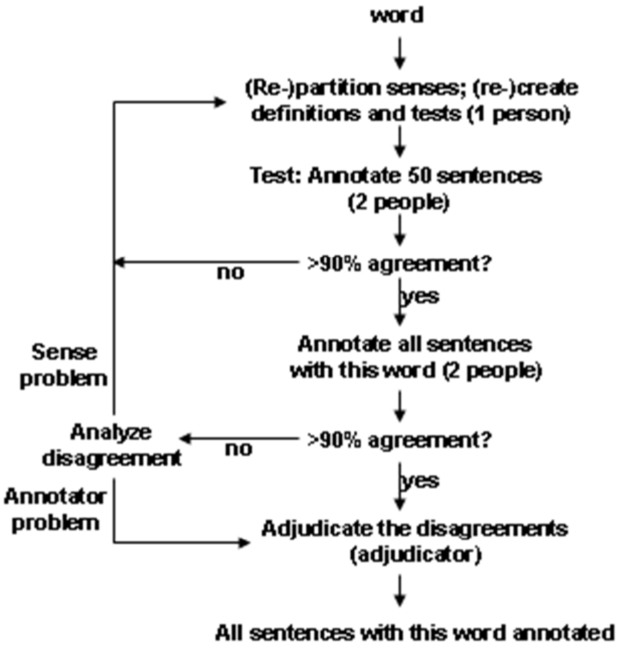
The English PropBank (http://verbs.colorado.edu/~mpalmer/projects/ace.html) has been developed at the University of Colorado under the supervision of Prof. Martha Palmer. The Chinese PropBank (http://verbs.colorado.edu/chinese/cpb) has been developed under the supervision of Prof. Nianwen Xue at Brandeis University and Prof. Martha Palmer at the University of Colorado. Arabic PropBank annotation has also been done under the supervision of Prof. Martha Palmer..

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 the following figure, 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.


word
(Re-)partition senses; (re-)create definitions and tests (1 person)
Test: Annotate 50 sentences (2 people)
>90% agreement?
no
yes
Annotate all sentences with this word (2 people)
Sense problem
Analyze disagreement
>90% agreement?
no
yes
Annotator problem
Adjudicate the disagreements (adjudicator)
All sentences with this word annotated

The word sense annotations for verbs was carried out at the University of Colorado, under the supervision of Prof. Martha Palmer, and the same for nouns was 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.

GI: operating or traveling via a vehicle NP (Agent) drive NP, NP drive PP	WN1: “Can you drive a truck?” WN2: “drive to school” WN3: “drive her to school” WN12: “this truck drives well” WN13: “he drives a taxi” WN14: “the car drove around the corner” WN:16: “drive the turnpike to work”
G2: force to a position or stance NP drive NP/PP/infinitival	WN4: “he drives me mad” WN5: “She is driven by her passion” WN6: “drive back the invaders” WN7: “she finally drove him to change jobs” WN15: “drive the herd” WN22: “drive the game”
G3: to exert energy on behalf of something NP drive NP/infinitival	WN11: “What are you driving at?” 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: excavate horizontally, as in mining	WN8: “Drive a nail into the wall” WN19: “drive a tunnel through the mountain”
G6: cause to function or operate	WN20: “steam drives the engine”
G7: search for a game, hunting	WN21: “drive the forest”

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 correctly 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 pseudo cleft 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

During the first three years of the GALE OntoNotes effort, we worked to link our word sense annotation to the Omega ontology (in particular, Omega 5). Each Ontology node represents a conceptualization. Word senses in OntoNotes are pooled into groups with (near-) identical meanings (similar to synsets in WorldNet), and these pools, treated as concepts, become ontology nodes in Omega. Each pool will be linked into Omega, allowing its parent and sibling nodes 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 information such as features, axioms to help in interpreting the entities and relations conveyed, domain tags, and *part-of* and other relations.

Omega 5 consists of two parts: an Upper Model of approximately 200 nodes and the ontology body. 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’. An automated 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.

We created and connected to the Upper Model approximately 2000 pools representing 5000 English noun and verb senses. In almost all cases, pools are connected directly to Upper Model leaf nodes; popular leaf nodes like *Artifact* govern several hundred pools, while many leaf nodes govern only one or two. This process was performed by between 4 and 7 poolers, operating independently. Only pooling or merging decisions exceeding a cutoff level of agreement have been included.

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 was 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 and Arabic, coreference annotation can be applied to a “*pro*” or

"*" 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.

While annotating the broadcast conversation data, we realized that the length of these documents, typically recordings of entire shows covering various topics, was prohibitive for full-document coreference annotation. We therefore chose to break the documents into multiple parts, breaking along story boundaries as much as possible, and to annotate coreference within those parts independently. The different parts of each document thus currently behave as independent documents, and the coreference chains do not carry any information across parts. This required some changes to the document format, as described in a later section. In the future, we hope to be able to fill in the coreference links that cross part boundaries, so as to create fully-coherent document-level annotation.

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 in a style similar 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 English Release Notes

3.1 English Corpora

The English OntoNotes corpus includes 300K of newswire 200K of broadcast news, 200K of broadcast conversation, and 145K of P2.5 data and 200K of Web data.

The newswire corpus 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. Parse and proposition data is also included for 546k of additional Wall St. Journal data from the Penn Treebank 2.

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.

The broadcast conversation data is 200K of LDC data, transcriptions of talk shows that include speaker and turn information as metadata. 100K of this data is parallel data, including 50K of English translated from Chinese and 50K of English translated into Chinese. For some of the parallel data (the Broadcast Conversation genre), tree-to-tree mapping information between the two versions is also provided. For other parallel data (the ECTB data), we only have document level parallelism.

The English web data is about 200K English out of which 55K of which is translated from Arabic and 75K translated from Chinese. There is 145K English P2.5 data, of which 80K is translated from Chinese and the rest is translated from Arabic. There is approximately 35K P2.5 data for each of Newswire, Broadcast News, Broadcast Conversation, and Web. There is also 85K of English web data consisting of single sentences selected to improve sense coverage

The telephone conversation corpus comprises about 100K of English CallHome data annotated with parse, proposition, name and coreference information

The pivot corpus comprises English translations of the 250K New Testament (NT) and a 100K portion of the Old Testament (OT). The NT is tagged with parse, proposition and coreference information whereas the OT is only treebanked.

3.2 English 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.

- english-translation-treebank-guidelines.pdf: This is a supplement to the other guidelines -- specifically dealing with parallel treebanks, webtext, etc.

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 3.4 below.

3.3 English 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 entire WSJ corpus, 85,348 propositions are annotated, covering almost all verbs[illegible]. The total number of verb types annotated is 3,101. The YR2 release involves the 200K English Broadcast News corpus, consisting of 33,800 propositions for 1,626 verb types which were double-annotated and adjudicated, including the “be” verb instances. 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: This release adds 1862 lemmas annotated from the Sinorama magazine, 888 lemmas from the Xinhua newswire and 1453 lemmas from the Broadcast conversation genre making a total of 59,013 propositions. The total number of frame files stand at 4072.

For a detailed description of the PropBank data fields and formats, see Section 6.5 below.

- english-propbank.pdf: English PropBank annotation guidelines

3.4 English Treebank/Propbank Merge Notes

In the initial Propbank 1.0, 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/PropBank 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.

3.4.1 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

[illegible] [illegible]

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.

3.4.2 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 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 co-indexed 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.

3.5 English Word Sense Notes

There are a total of 264,622 words in the combined NW, BN, BC, WB and Sinorama corpora tagged with word sense information. These cover 1,338 noun and 2,011 verb types. A total of 6,147 word senses have been pooled and connected to the Ontology in about 2,732 pools.

Our internal quality control tests are applied separately to the instances of each word in each corpora. Thus words can have adjudicated word sense data in some of the corpora but not in others.

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 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.

3.6 English Coreference Notes

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

Coreference coverage of the ECTB 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. In addition to that, there is a file called “single-annotated-coref-files.txt” in the metadata directory which contains a list of all files that has only been singly annotated.

3.7 English 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.

4 Chinese Release Notes

4.1 Chinese Corpora

The Chinese portion of OntoNotes 5.0 includes 250K words of newswire data, 270K words of broadcast news, and 170K of broadcast conversation.

The newswire data 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 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.

The broadcast conversation data is 170K words, taken from LDC's GALE data. 50K of the originally-Chinese data has also be annotated in English, and another 55K of the Chinese data represents translations into Chinese from originally-English broadcast conversations.

The Web data includes 215K tokens of which 15K are from the P2.5 evaluation and 86K are from the Dev09 data. Futher, the 110K of Web data consists of 40K parallel Chinese origin data and 70K parallel English origin data.

The telephone conversation corpus comprises about 100K of Chinese CallHome data annotated with parse, proposition, name and coreference information

4.2 Chinese Treebank Notes

The annotation of syntactic structure trees for our Chinese newswire data was taken from the Chinese Treebank 5.0 and updated with some corrections. Some of the 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 remaining Chinese corpora 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.
- chinese-treebank-parses-bc-addendum.txt: Mentions a couple more tags that have been added to the Chinese treebank to deal with Broadcast Conversation data.

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)

4.3 Chinese 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 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 ”Adding Semantic Roles to the Chinese Treebank” by Xue and Palmer, (*Natural Language Engineering*, to appear).

4.4 Chinese Word Sense Notes

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 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.

4.5 Chinese Coreference Notes

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

Adjudicated coreference coverage of the broadcast news portion is not 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.

4.6 Chinese 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.

5 Arabic Release Notes

5.1 Arabic Corpora

The Arabic portion of OntoNotes 5.0 includes 300K words of newswire data. It also comprises parse and proposition layers on 200K more words.

The newswire data is taken from the 400K Arabic Treebank Part 3. V3.1 (ATB P3 V3.1)

OntoNotes Arabic annotation began using the version 2.0 of the Arabic Treebank (ATB P3 V2.0), During the GALE program, it was decided to revise the Arabic Treebank to make it more consistent. We stopped annotating propositions, but continued to annotate word sense and coreference data on the existing trees, since we anticipated that the changes made in noun and verb parts of speech and in the NP structures would generally be localized enough that we could automatically map our existing annotation onto the new structures. When ATB P3 V3.1 was released, we wrote procedures to map our existing annotations over to the LDC's revised trees. However there are potential cases where the annotation will need to be revised, possibly owing to new NPs introduced in the revised Treebank, or to changes in tokenizations that would expose new verb and noun tokens.

5.2 Arabic Treebank Notes

The Arabic parse trees in OntoNotes come from the LDC's Arabic Treebank effort. (OntoNotes was not funded to do its own Arabic Treebanking.) The trees in this release are taken from their Arabic Treebank Part 3 v 3.1 release (LDC2008E22[illegible]). (This August 2008 release updated the earlier LDC2005T20.)

Owing to complex morphology in Arabic, each token in the Treebank has an associated morphological analysis that is distributed with the trees. We have included this information in the OntoNotes release as .lemma files. The format of this file is as follows:

[illegible]

[illegible]

This information can be used to get morphological information for each token in the Treebank. Part of the information in the .lemma file pointer to offsets in the source files that were used for Tree banking. These .sgm files are also part of the release.

We have copied the relevant readme files from the Arabic Treebank release into the documentation folder which should provide information in interpreting the data. There is also a mapping from Arabic parts of speech into Penn Treebank parts of speech which is available as metadata in the Arabic metadata folder as "mapping-to-PennPOS-tags.lisp" Further information about the modifications to the original Arabic Treebank guidelines can be found in the following paper:

Mohamed Maamouri, Ann Bies, Seth Kulick . Enhancing the Arabic Treebank: A Collaborative Effort toward New Annotation Guidelines. 2008. *In Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008)*, Marrakech, Morocco, May 28-30, 2008. A copy of the paper is included in the documents directory.

5.3 Arabic Word Sense Notes

The word sense annotation for the Arabic data is closely following the pattern set for English. 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.

5.4 Arabic Coreference Notes

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

5.5 Arabic Name Annotation Notes

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

.

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.

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 Arabic 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 (ontonotes-v5.0.sql.gz), 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
- names
- lemma

The .lemma files are specifically for arabic treebank. These contain lemma information for each token in the treebank – the vocalized, unvocalized and surface forms, along with lemma information. These are redistributed version of what are .pos files in the Arabic Treebank.

In addition to these starting from OntoNotes version 3.0 there are two more files:

- speaker
- parallel

The .speaker files contain information on speakers for conversational genres and the New Testament subcorpus`, and the .parallel files contain either document level or sentence/tree level mapping between files that constitute parallel data.

The input and output versions of the annotation files are substantially identical in content, differing only in formatting.

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.

For consistency and convenience of naming, we had to map some originally different and usually long filenames into a shorter, consistent form. Under each genre directory there is a file called [illegible] which contains the mapping from the existing files to their original counterparts – which are most likely the LDC filenames. Users who would like to get to know the original names of the files in an LDC catalog, might want to use this mapping file. We were careful not to change already existing mapping such as one for the WSJ data, and the ECTB and CHTB data. We do however provide the original filenames for the same data. In addition to providing this mapping, the [illegible] file also contains path to files that were used to extract sentences from – in case where we only have annotations on selected sentences from a document along with the space separated token offset that would be useful to identify the word that was annotated. The full files provide more context for the annotation.

Following is a small excerpt from the [illegible] file:

```
[illegible]
[illegible]
[illegible]
[illegible]
[illegible]
[illegible]
[illegible]
[illegible]
[illegible]
..
..
[illegible]
[illegible]
[illegible]
[illegible]
[illegible]
[illegible]
```

We have also tried to provide original source ([illegible]) files wherever possible. The correspondence between the source files and the [illegible] files is not necessarily 100% as some edits had to be made during treebanking.

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

The OntoNotes database schema is shown in Figure 1. Owing to space constraints we have not displayed all of the tables in the figure. More detailed information is present in the accompanying API document.

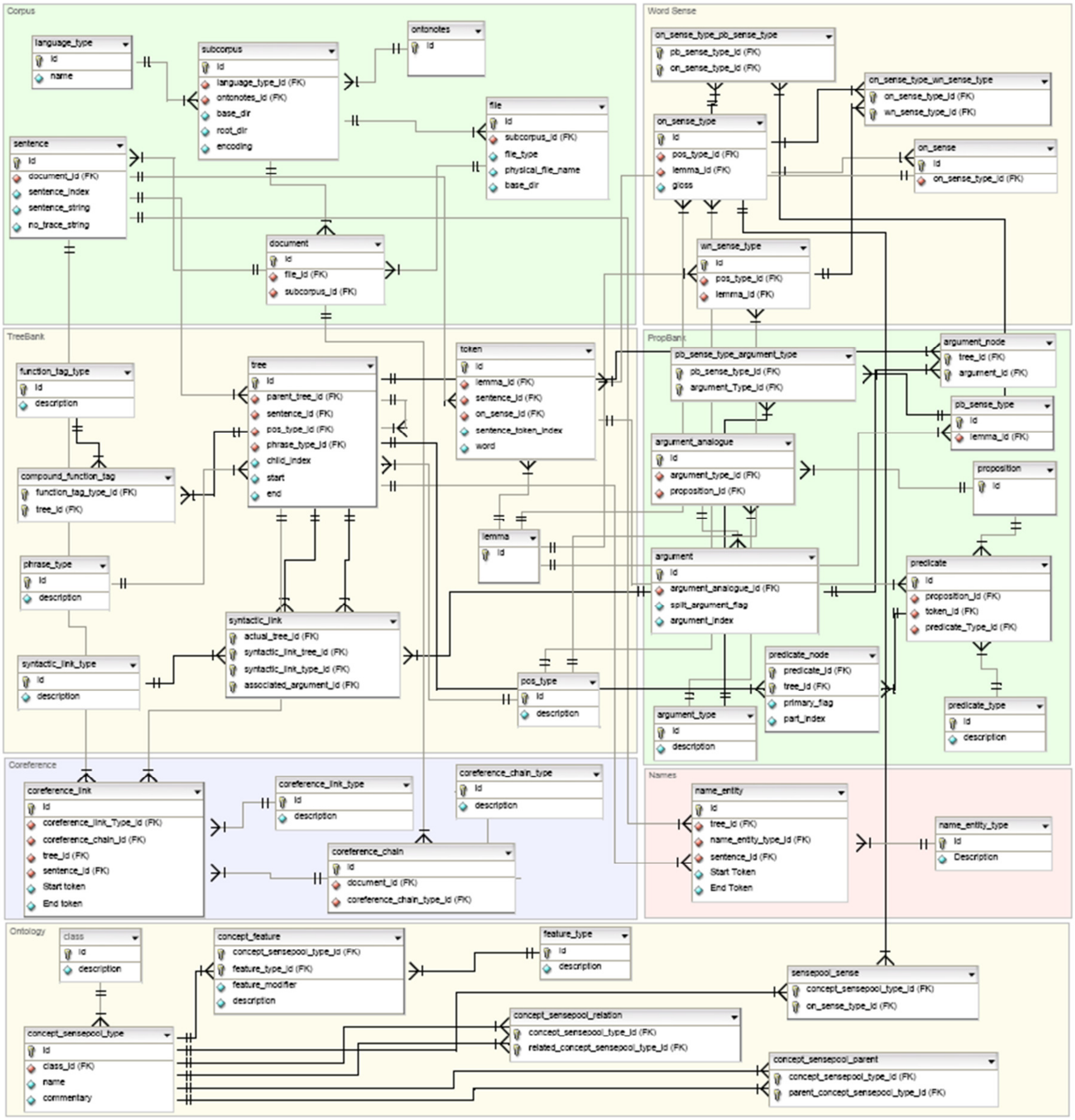
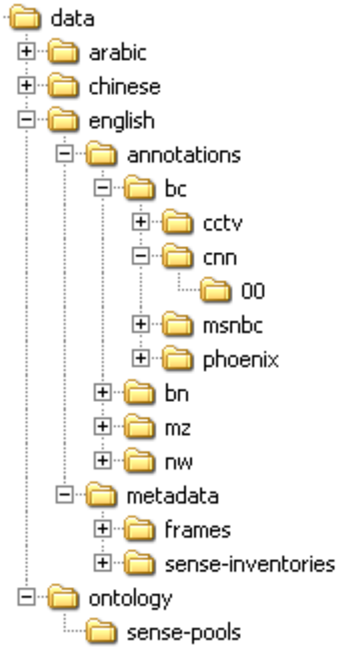


Figure 1: The OntoNotes Database Schema

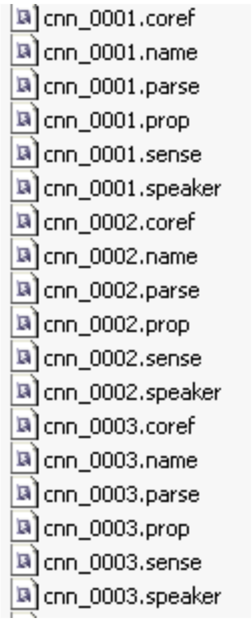
The database tables are shown divided into six logical blocks, with one block for the textual corpus data, and then a block for each type of semantic annotation: Treebank, Proposition Bank, Word Sense, Coreference, and Name Entities. Each of the annotation types involves adding additional meta information 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.



```
data
+ arabic
+ chinese
- english
  - annotations
    - bc
      + cctv
      - cnn
          00
      + msnbc
      + phoenix
    + bn
    + mz
    + nw
  - metadata
    + frames
    + sense-inventories
- ontology
    sense-pools
```

[illegible]



```
cnn_0001.coref
cnn_0001.name
cnn_0001.parse
cnn_0001.prop
cnn_0001.sense
cnn_0001.speaker
cnn_0002.coref
cnn_0002.name
cnn_0002.parse
cnn_0002.prop
cnn_0002.sense
cnn_0002.speaker
cnn_0003.coref
cnn_0003.name
cnn_0003.parse
cnn_0003.prop
cnn_0003.sense
cnn_0003.speaker
```

[illegible]

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 may seem a bit redundant, but it 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. 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. Barring a relatively small number of cases, most coreference links align with the tree nodes. This is less so in case of name entities. The fact that a name or coreference span does not align with the tree is represented by a preceding exclamation point (as for the name spanning tokens 2 and 3 in the example OntoNotes Normal Form below):

[illegible]

[illegible]

[illegible]

```
               [illegible]   [illegible]
               [illegible]   [illegible]
               [illegible]   [illegible]
               [illegible]   [illegible]
               [illegible]   [illegible]
               [illegible]   [illegible]
               [illegible]   [illegible]
               [illegible]   [illegible]
               [illegible]   [illegible]
               [illegible]   [illegible]
               [illegible]   [illegible]
               [illegible]   [illegible]
               [illegible]   [illegible]
               [illegible]   [illegible]
               [illegible]   [illegible]
               [illegible]   [illegible]
               [illegible]   [illegible]
               [illegible]   [illegible]
               [illegible]   [illegible]
               [illegible]   [illegible]
               [illegible]   [illegible]
               [illegible]   [illegible]

    [illegible]
               [illegible]   [illegible]
               [illegible]   [illegible]

    [illegible]
               [illegible]   [illegible]
               [illegible]   [illegible]
               [illegible]   [illegible]
               [illegible]   [illegible]
               [illegible]   [illegible]
               [illegible]   [illegible]
               [illegible]   [illegible]
               [illegible]   [illegible]
               [illegible]   [illegible]
               [illegible]   [illegible]
               [illegible]   [illegible]
               [illegible]   [illegible]
               [illegible]   [illegible]
               [illegible]   [illegible]
               [illegible]   [illegible]
....
....
....

------------------------------------------------------------------------
```

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 word sense, 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.

```
[illegible]
```

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:

filename sentence terminal tagger frameset ------- proplabel proplabel

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

- filename: the name of the parse file in English or Chinese .
- 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:
[illegible]
the terminal numbers are: John 0; wants 1; *-1 2; to 3; swim 4
In the Chinese example:
[illegible] 货币 [illegible] 回笼 [illegible] 的 [illegible] 增加 [illegible]
, [illegible]
[illegible] 为 [illegible] 平抑 [illegible]
[illegible] 全 [illegible]
[illegible] 区 [illegible] 物价 [illegible] 发挥 [illegible] 了 [illegible] 作用
[illegible] 。[illegible]
the terminal numbers are:
货币 [illegible] 回笼 [illegible] 的 [illegible] 增加 [illegible] , [illegible] 为 [illegible] 平抑 [illegible] 全 [illegible] 区 [illegible] 物价 [illegible] 发挥 [illegible] 了 [illegible] 作用 [illegible] 。[illegible]

- 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).
- 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: [illegible]
  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
  [illegible]
  A syntactic relation of "[illegible]" represents the NP immediately dominating the terminal [illegible] and a syntactic relation of "0:2" represents the "S" node.

  In the Chinese sentence
  [illegible] 这些 [illegible] 已 [illegible] 开业 [illegible] 的

[illegible] 外商 [illegible] 投资 [illegible] 企业 [illegible] 绝大部分 [illegible]
[illegible] 生产 [illegible] 经营 [illegible] 状况 [illegible] 较 [illegible] 好 [illegible] 。[illegible]
the address of "[illegible]" represents the top IP node and [illegible] represents the CP node

- form 2[illegible]
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
[illegible] 这些 [illegible] 已 [illegible] 开业 [illegible] 的 [illegible] 外商 [illegible] 投资 [illegible] 企业 [illegible] 绝大部分 [illegible] 生产 [illegible] 经营 [illegible] 状况 [illegible] 较 [illegible] 好 [illegible] 。[illegible]
the address of of "[illegible]" represents the fact nodes [illegible], [illegible] and [illegible] 外商 [illegible] 投资 [illegible] 企业 [illegible] are coreferential.

- form 3: [illegible]
A split argument, where there is no single node that captures the argument and the components are not coreferential, e.g. 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
[illegible]
The phrasal verb "keep_on" would be identified with the syntactic relation [illegible].

- form 4: [illegible]...
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
[illegible]
The proplabel [illegible] 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[illegible]. The interpretation of this argument is that the "causer of action" ([illegible] of have.04) is signified by the following trace-chain: *T*-4 --> that --> ([a series][of intrigues])

- form 5: [illegible]
This indicates that either of the two nodes represent an ICH node.

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:

- ■■■ - possessor
- ■■■ - possessee
- ■■■ - coordinator
- ■■■ - predicate
- ■■■ - 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:

- ■■■ - adverbial, default tag
- ■■■ - beneficiary
- ■■■ - conditional
- ■■■ - directional
- ■■■ - discourse
- ■■■ - degree
- ■■■ - extent

- [illegible] - frequency
- [illegible] - location
- [illegible] - manner
- [illegible] - negation**
- [illegible] - purpose and reason
- [illegible] - temporal
- [illegible] - topic

In addition to these, a new [illegible] tag has been added to the PropBank. Two LINK tags are introduced to capture two types of connections that provide useful information which is not grounded in syntax and is therefore absent from the Treebank. These are [illegible] (“SeLectional Constraint”) and [illegible] (“Pragmatic CoReference”). In the original PropBank v1.0 release, this distinction was not made, and the nodes were associated with the respective argument.

Let’s consider an example with a [illegible].

```
[illegible]
```

Here, [illegible] links [illegible] to the constituent represented by node [illegible] in the parse tree. This link requires one common node shared between the [illegible] and the node [illegible] (which represents [illegible]). Let's call this the anchor node. The link type is [illegible] or [illegible] depending on the type and syntactic properties of these nodes. For more detailed examples, the reader can refer to the PropBank guidelines addendum document ([illegible]) included in the release. Roughly speaking, relative clauses and infinitival clauses would be tagged as [illegible], and reduced relative clauses will be tagged as [illegible]

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

```
[illegible]
[illegible]
[illegible]
[illegible]
[illegible]
[illegible]
[illegible]
[illegible]
[illegible]
....
....
....
```

Starting from release 5.0, we have modified the PropBank view to eliminate information that is duplicated in the tree. For example previously the annotation for the arguments of a predicate contained node IDs for both nodes that were either traces, or co-indexed with trace nodes, and represented some argument of a predicate. Now, only one of the two would be present in the PropBank view for that argument. This was done to prevent inconsistent annotation with respect to the trees.

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. These lines can be of one of two formats.

```
[illegible]

[illegible]

[illegible]
```

The form in [illegible] indicates that the instance has been adjudicated. The “[illegible]” placeholders fill slots in the format that was used internally by the word sense annotation tool. Form [illegible] indicates that the instance was either single or double annotated. In case it was double annotated – both the annotators had selected the same sense. All double-annotated disagreements were either adjudicated or removed from the data.

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

```
[illegible]
[illegible]
[illegible]
[illegible]
[illegible]
[illegible]
[illegible]
[illegible]
[illegible]
[illegible]
....
....
```

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.

Beginning in version 3.0, owing to the fact that we had to split very long broadcast conversation documents into multiple parts for coreference annotation, we have changed the format of the coreference document to make the DOCNO an attribute to the DOC tag, and add a PARTNO attribute to the TEXT tag to identify the part in the coreference document. Since coreference chains do not cross part boundaries, we append the coreference link ids with the part number to avoid potential confusion.

The Arabic data is distributed in buckwalter format which contains some non-SGML friendly tags like “>”, “<”, “&” which we have replaced with “-LAB-“, “-RAB-“ and “-AMP-“ respectively similar to the way parentheses were dealt with in the Penn Treebank.

Beginning in version 5.0, we have added a SPEAKER attribute to each coreference link that is coreferent with a particular speaker in case where this data is available. In the Web data, SPEAKER can represent the author of the comment or blog, newsgroup, etc.

Another modification do the data was in regards to the way subtoken annotation was dealt with. Previously entities that were part of a word (defined as sequence of non-space characters) and which were not spanning the entire treebank token – such as anti-Wal-Mart would be expanded to include the entire token. This was not quite optimal, so we modified the coreference markup to include two new attributes S_OFF and E_OFF representing start and end offsets into the token respectively. These are both inclusive. Unless specified either of the two attribute values should be assumed to be 0. Usually only one of S_OFF or E_OFF is required when there exists a subtoken annotation as the character span representing an entity either align with the left or the right word boundary. In some cases however, we have to specify both offsets.

[illegible]

6.8 Entity Names View

The entity names annotation view is formatted using in-line ENAMEX markup. We also extended the representation to allow subtoken annotations as in the coreference data.

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

```
[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

...
...

[illegible]
```

6.9 Parallel View

For data that is parallel between two languages, the source language as well as the target language documents have a associated .parallel file which contains mapping information. An example from the source is mentioned below. The target side file just points to the document that it is a translation of along with the tree-to-tree mapping when available as follows:

Original Document: [illegible]

```
[illegible]
```

Translated Document: [illegible]

```
[illegible]
[illegible]
[illegible]
[illegible]
[illegible]
[illegible]
[illegible]
....
....
```

Note that the ids of the two documents in both languages don't match. Also, although not clear in this example, the second and third columns represent the tree index in the original and the translated document respectively (although the .parallel file that contains the mapping accompanies the translated document)

6.10 Speaker View

For the Broadcast conversation data, there is a .speaker file associated with each document which contains speaker metadata such as turn start/end time, gender, nativity, etc.

Excerpt from a speaker file is shown below:

Document: [illegible]

```
[illegible]
[illegible]
[illegible]
[illegible]
[illegible]
[illegible]
[illegible]
```

Since the .speaker file has space separated values in each row, and some speaker names have spaces in them, we have converted all the spaces in speaker names to an underscore as seen in the example above.

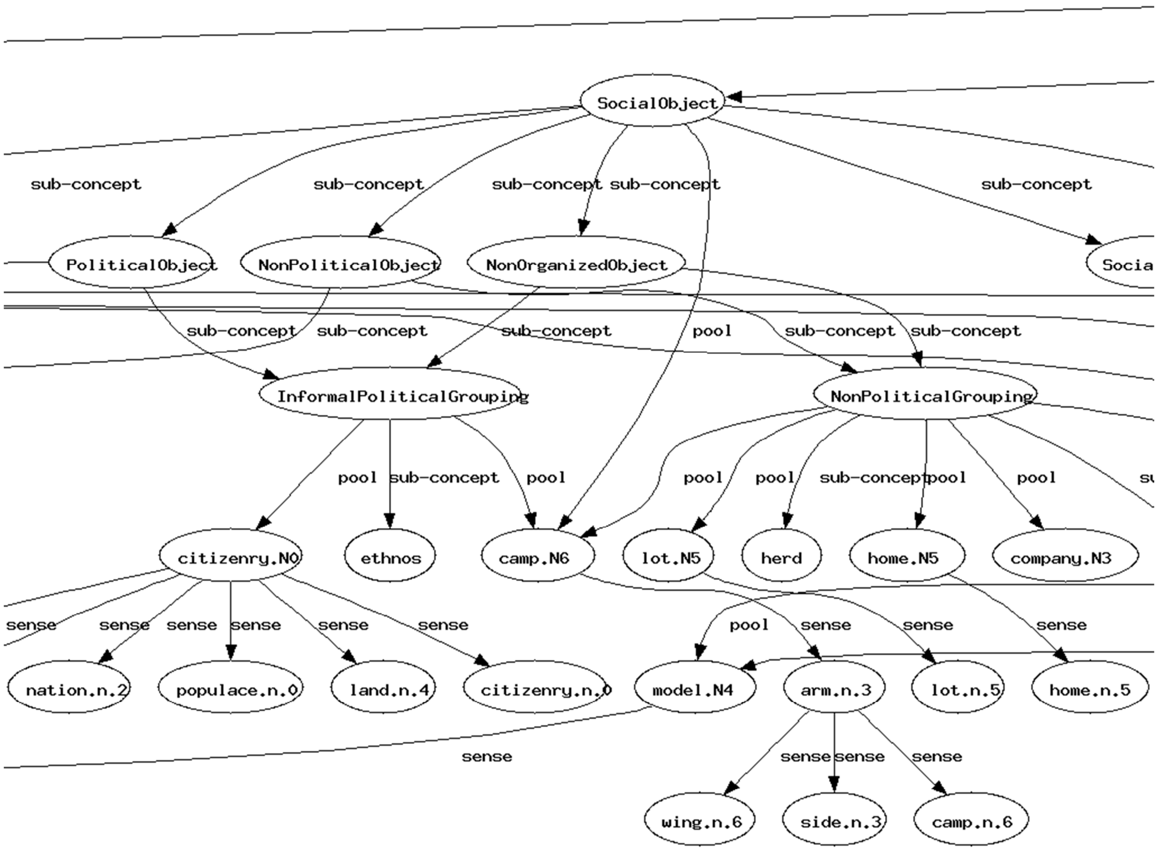
6.11 Ontology View

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:

```
[illegible]
    [illegible]
    [illegible]
    [illegible]
                    [illegible]
    [illegible]
    [illegible]
    [illegible]
    [illegible]
    [illegible]
    [illegible]
                           [illegible]
    [illegible]
    [illegible]
    [illegible]
    [illegible]
    [illegible]
    [illegible]
    [illegible]
    ...
[illegible]
```

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.



6.12 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.

6.12.1 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 Arabic and English directories containing the frames files for each of the verbs covered.

6.12.2 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 PropBank frame, as appropriate.

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

6.13 Access Script Documentation

The database contains the merged layers of annotation for both the English and Arabic 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 34■ 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.

Weischedel, R., Hovy, E., Marcus, M., Palmer M., Belvin, R., Pradhan, S., Ramshaw, L., Xue, N. OntoNotes: A Large Training Corpus for Enhanced Processing. In *Handbook of Natural Language Processing and Machine Translation: DARPA Global Autonomous Language Exploitation, Eds. Joseph Olive, Caitlin Christianson, and John McCary, Springer, 2011.*