BULB: A Unified Lexical Browser

Catherine Havasi*, James Pustejovsky*, Marc Verhagen*

*Brandeis University
Computer Science MS018
415 South St.
Waltham, MA, 02454
{havasi, jamesp, marc}@cs.brandeis.edu

Abstract

Natural language processing researchers currently have access to a wealth of information about words and word senses. This presents problems as well as resources, as it is often difficult to search through and coordinate lexical information across various data sources. We have approached this problem by creating a shared environment for various lexical resources. This browser, BULB (Brandeis Unified Lexical Browser) and its accompanying front-end provides the NLP researcher with a coordinated display from many of the available lexical resources, focusing, in particular, on a newly developed lexical database, the Brandeis Semantic Ontology (BSO). BULB is a module-based browser focusing on the interaction and display of modules from existing NLP tools. We discuss the BSO, PropBank, FrameNet, WordNet, and CQP, as well as other modules which will extend the system. We then outline future extensions to this work and present a release schedule for BULB.

1. Introduction

Natural language processing researchers today have access to a wealth of information about words and word senses. Such information is of critical importance to a wide variety of researchers: from the researcher trying to build a large-scale ontological system to the theorist exploring a small part of the language in detail.

Besides a variety of machine readable dictionaries, the major lexical resources currently available include: WordNet (Fellbaum, 1998), PropBank (Kingsbury and Palmer, 2002), Corpus Pattern Analysis (CPA) (Pustejovsky et al., 2004), FrameNet (Fillmore and Baker, 2000), EuroWordNet (Vossen, 1998), SIMPLE (Busa et al., 2001), and a variety of "word in context" tools (concordance tools) such as CQP (Christ et al., 1991) and BONITO (Rychly and Smrz, 2004). These resources and their accompanying browsers provide ample opportunities for research. However, this vast quantity of information can be stifling as one deals with the many wrappers and browsers required to access what is useful in this information. A more compact and seamless tool to access this collection of information would be of great value to the NLP community.

1.1. Goal of the project

To fulfill this need, we have developed a unified lexical palette called the Brandeis Unified Lexical Browser (BULB). Our aim for this browser and its accompanying back-end is to provide the NLP researcher with a coordinated display from many of the available lexical tools, focusing on a newly developed lexical database, the Brandeis Semantic Ontology (BSO) (Pustejovsky et al., 2005). We intend to provide coordinated access to information from the BSO, Propbank, WordNet, and a concordance tool by default.

We plan on having two front-ends for BULB: an online browser and a site-based browser which would be more customizable for the user's needs. The online browser will also be customizable, as well as expandable by outside parties. This will allow researchers to use the tool's interface to build a custom browser to fit their needs. Lexical resource developers can provide BULB modules which the extended NLP community can use to access their emerging resource. The individual NLP researcher can customize which modules he wishes to use to fit her interests and current projects. A project can customize a cite-specific BULB system to facilitate development work on a corpus or other major project.

The browser interface and its modules will be described in additional detail later in this document.

The online browser will be available for free public use starting in the summer of 2006, following local user testing. The site-based browser will be available at a later date.

2. Functionality

The main function of this environment is to provide a unified front end for a researcher who wishes to access information about a given word, and to provide a platform for browsing and editing ontologies. There are many popular lexicons and ontological systems currently available to be downloaded, and each provides different sorts of information about a word. Our goal is to combine the most popular and diverse of these information sources, and provide the researcher with coordinated results from these tools. The researcher will no longer be impeded by the cumbersome use of multiple browsers for each information source they wish to incorporate. There is a base set of modules, each of which incorporates a tool. Additionally, the user will be able to install additional modules for extending the system to be more tailored to the individual user's needs. BULB also aims to serve as an authoring tool in the development of the BSO.

2.1. Interface

We are currently planning to develop two BULB frontends. The first is an online tool which users can access and log into from any environment. The second is a more customizable stand-alone, local, GUI interface. The interface to BULB, in either software form, is a series of modules

Module	Tool	Domain	Release
BSO	BSO	All	Early Web
PropBank	PropBank	Lexicon Verbs	Early Web
BSO Tree	BSO	All	Early Web
PropBank	WordNet	Lexicon items	Early Web
CQP	CQP/BNC	Lexicon items	Early Web
NomBank	NomBank	Lexicon Nouns	Full Web
CPA	CPA	Some lexicon verbs	Full Web
Bonito	Bonito/BNC	Lexicon items	Full Web
Sketch Engine	Bonito/BNC	Lexicon items	Full Web
Module Creation	—	All	Full Web
Editing	—	All	Standalone

Table 1: BULB Modules

which contain the information provided by each of the selected lexical tools for a given query. Each of the modules is presented in its own tab, and the user can switch between tabs to view additional information. An example of this layout is shown in 2.1..

The most important tab is the "Overview" tab, which is the first tab a user sees upon the completion of a query. The overview tab contains a BSO sense for the word, along with a sampling of the results for the given word from three or four of the installed modules, as configured by the user. In the web-based version of BULB, users must register for free before using the software; this allows them to customize their BULB configuration, including which modules are loaded and which appear in the overview screen. An example of the overview tab in the Web-based BULB is shown in 2.1.

To search for a word, the user simply types the word into the search box, as shown in 2.1.. The user may search for a word or combination of words as part of the "Lexical" search. He may also choose a part of speech to further restrict his search. In the "Ontological" search, the user may search for an item which is part of the BSO type system, described in 2.3.1.

2.2. Module Plug-in Design

The BULB system is made to be easily extendable by members of the Brandeis Laboratory for Linguistics and Computation and by other researchers in the field of NLP. We will be releasing a module authoring toolkit to allow others to easily make modules for the browser to fit their particular NLP system or their research interest. We encourage groups to create modules for the use of others to further extend the collaborative spirit of BULB.

The module authoring toolkit would contain instructions on how to create a module which we could "plug in" to the BULB browser. Modules would be written in Python and make use of the existing BULB user interface, and would be easy for the BULB staff to install. The system is designed so that a simple module would only need to be placed in a module library directory to be integrated into the system. This would allow creators of lexical resources to incorporate their research into BULB and would allow the number of available BULB modules to grow quickly.

2.3. The Modules

We are currently including several modules into the BULB browser by default: the Brandeis Semantic Ontology, WordNet, and Propbank, as well as concordance information. The BSO is described below. The WordNet module provides relational information and synonym-set information for a given word. The Propbank module provides lexical selection information and argument role information for a given word, while the concordance tool provides sample usage and usage in context. We will add additional modules for the other lexical resources listed above. These modules will be downloadable by the user and will plug into the system allowing a more customizable browsing experience. A list of currently planned BULB modules can be seen in 2..

2.3.1. The BSO

The BSO is a new lexically-based ontology in the tradition of Generative Lexicon (Pustejovsky, 2001; Pustejovsky, 1998). In particular, it focuses on contextualizing the meanings of words through a rich system of types, including qualia structures. For example, if one were to look up the phrase RED WINE in the BSO, one would find its type is WINE and its type's type is ALCOHOLIC BEVERAGE. An example of a qualia structure is shown in 3. A user can thus use the BSO to find out where in the ontological type system WINE is located, what RED WINE's lexical neighbors are, and its full set of part-of-speech and grammatical attributes. Other words have a different configuration of annotated attributes depending on the type of the word and other factors.

The first of two BSO BULB modules displays the various lexical senses for the word and other corresponding information. This includes the qualia structure, the sense's ontological type, grammatical information, and the types of its arguments. One can use this module to specify a particular sense to display in the other BULB modules, which also sets the part of speech of the search to be the part of speech of that BSO sense.

Currently, we cannot match senses across different non-BSO modules. Ontological unification is a difficult and much studied problem, and in the future we feel it would be interesting to attempt to match senses across various BULB modules, though this is outside the scope of this paper.

The second BSO BULB module is the "BSO Tree", which

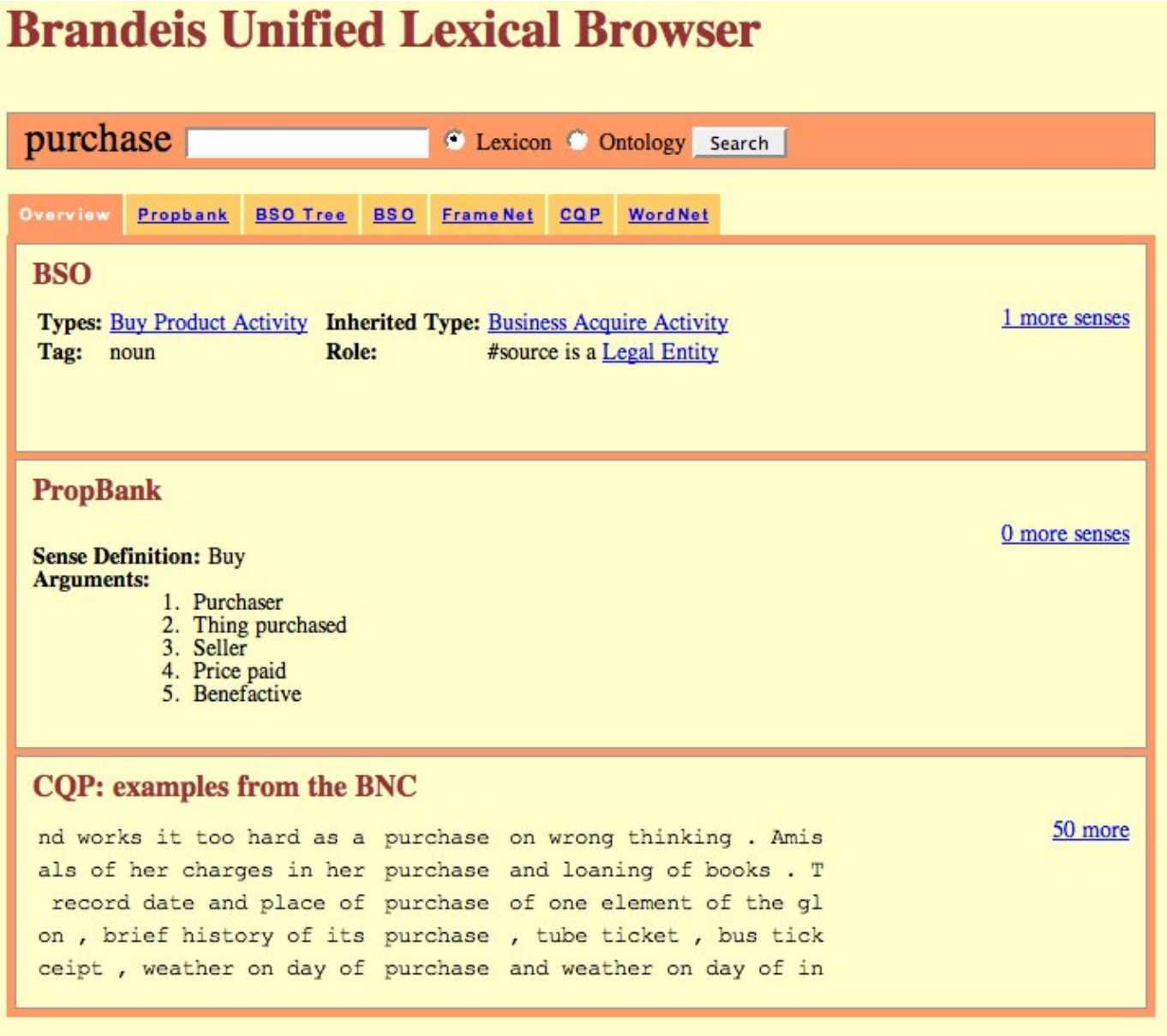

Figure 1: The Brandeis Unified Lexical Browser.

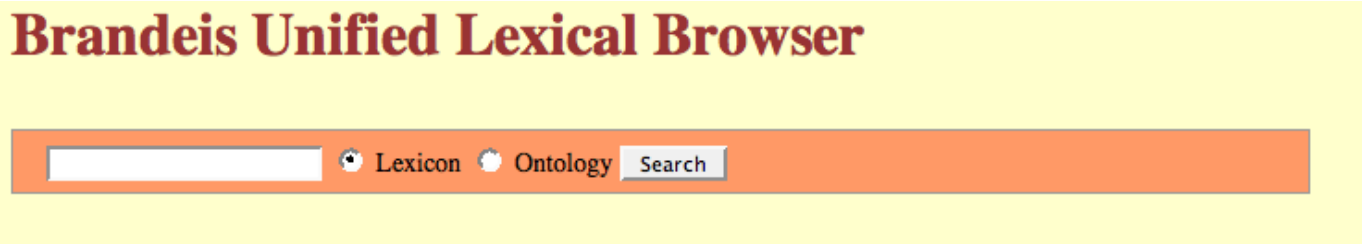

Figure 2: The search box for the BULB.

shows the tree of ontological types in the area surrounding the currently selected BSO sense. Here, one can view the ontological parents and children of a word's type, as well as any sibling words which have the same type as the given word.

2.3.2. PropBank

Another of our prominent modules is the PropBank module, which has additional information for lexical verb searches. This module is based on a locally-developed Python wrapper and a copy of PropBank hosted locally.

The PropBank module displays the available PropBank

Figure 3: A BSO Qualia Structure

Indirect Telic: DRINK ACTIVITY
Indirect Agentive: MAKE ALCOHOLIC BEVERAGE ACTIVITY
Constitutive: ALCOHOL
Has Element: ALCOHOL
Made Of: GRAPES

senses of the word, along with the definition and arguments for each sense. If possible, the arguments are linked back into BULB so that clicking on an argument brings up a search for that word.

2.3.3. WordNet

The WordNet module displays some of the more general WordNet information about a search target. For each applicable sense in WordNet, the module displays the part of speech and definition. It also displays the *synset* information, such as hypernyms and hyponyms. The module also displays example sentences for each word sense.

This module is linked to the rest of BULB in a similar manner to the PropBank module. Although we cannot yet match senses between tools, we provide links from the synset information to their corresponding BULB searches to aid browsing.

2.3.4. FrameNet

The FrameNet module maps a word to the possible frames for that word. Similarly to the way the BSO module displays information about the parent ontological type for a word, the FrameNet module displays information about the corresponding frame.

Also, as in the BSO module, the user can choose to "focus" on a frame. This currently only makes the other modules aware of word and part of speech for the given selection. As discussed earlier, building a better link between the frame portion of FrameNet and the other BULB ontological tools is a difficult and interesting problem.

2.3.5. Concordance Tool

We are currently using the CQP concordance tool, though we plan to create a BONITO module shortly. The CQP concordance tool currently displays sentences from the BNC which contain the target word. For the overview screen, it displays 15 concordance matches and for its module, it displays 50. One can further refine this query by choosing the part of speech of the target word.

We would also like to create a module for the BONITO concordance system. This module would contain concordance information as well as information from BONITO's word sketch engine and thesaurus.

2.3.6. Other Modules

Modules for the BULB browser that are currently under development include modules for NomBank (Meyers et al., 2004) and CPA. The CPA module would allow access to information for the verbs which have been completed in the current version of CPA. Also, to complement our FrameNet module, we would like to add a NomBank module which would display the argument sets for nouns. Ideally, a noun target word search would bring up a NomBank module while a verb target word search would bring up a PropBank module instance. A EuroWordNet module is also under consideration.

Also, we plan on allowing others to create modules for the BULB system through the expansible module system, which can all be accessed through the site-specific browser. In the online browser, we plan on including a selection of modules which would be useful to a variety of researchers.

3. The Brandeis Unified Lexical Browser

The current browser focuses on two media for the deployment of the BULB platform. The first is an extensible stand-alone GUI browser and authoring tool. The second is a simpler online browser with access to a few of the included ontologies.

The web-based browser is platform and location independent, and is viewable from anywhere a researcher can connect to the internet, once the user has obtained a BULB login. The web browser, however simple, does allow for user customization. A user of the web-based interface will be able to choose from a wide variety of BULB modules to customize his individual BULB browser for his research needs.

The GUI browser is a more complex version of the its web-based relative. BULB focuses on a word sense's entry in the BSO and displays information about that word sense which can be found in other lexical resources. The interface is transparent, blending the ontologies together into a palette which is easy for a researcher to work with. Also, users will be given the option to edit a local copy of some parts of the BSO using this browser to further tailor the database to their needs. Members of the BSO research group also plan on using an expanded version of the authoring tool to continue developing the BSO type system.

3.1. Evaluation Phase

We plan to do a formal user study of the BULB interface both from a UI standpoint as well as from the standpoint of a natural language processing researcher. We would like to make sure that the interface is as useable as possible for people approaching BULB from various different research perspectives.

A usability study will be done in April to evaluate the interface. We would like it to be simple to navigate between the modules and to customize the search functionality. We have also been receiving feedback from a set of test NLP researchers who are currently using BULB as a reference tool.

3.2. Demonstration at LREC

By May, we expect to have an integrated lexical palette available to researchers for download and integration. The proposed demo will include the online BULB browser and associated modules. We will also be demoing the module "plug-in" system as well as modules created with this system. We will also have an initial version of the stand-alone browser and editing tool available. We will demonstrate the tool and the interaction between the various resources as well as the addition of various modules. We will show how to log-in and customize the browser for a user's individual needs.

We plan to display the BSO, BSO Tree, PropBank, WordNet, FrameNet, CPA, CQP and NomBank modules. Also, we plan to demo other optional modules.

3.3. Release Plan

We plan to be releasing both these browsers over the course of the next six months. The first release will be the online web-based browser system. We plan to release a simple version of this tool at the LREC 2006 conference in May. During the summer we will release the more complex full online browser. This browser will allow logins and customizations as well as containing several more optional modules. When we release the full online module, we will also release the module authoring toolkit to give others an opportunity to begin create modules for the BULB platform. Next, we will begin welcoming submissions of additional modules, which will be distributed under the same license as the BULB browser.

We plan to release the stand-alone site-specific browser/editor tool near the release time for the BSO ontology, which is currently projected to be in the late summer or early fall of 2006.

Both browsers will be released under Creative Commons licenses for the research community, with the stipulation that our browser is not included in a commercial product.

3.4. The Editing Tool

Although we will not discuss the editor in detail here, we will note a few facts about it. The editor portion of the stand-alone BULB browser is designed to allow an ontological developer to access relevant facts about an entry from other NLP tools.

The browser-editor, which will be released, will allow access to a site-specific copy of the BSO database for editing. Currently, designs support the editing of the lower levels of the type system and the lexical entry system.

The editor will allow a user to edit information about a lexicon entry such as grammatical and argument information. The user will also be able to move a lexical item to a new place in the type ontology structure of the user's custom BSO or to create a new lexicon item.

The user can also edit the lower and middle level type system: for example, changing the "Direct Telic" qualia for a given type. The editor tool will display to the end user the effect his changes will have on other parts of the database and maintain the consistency of the database through the editing process.

A system of locks allows only one user to edit a given portion of the BSO at a given time. This, along with a series of consistency checks that are run on each edit, enables us to maintain internal consistency of the database.

A more powerful version of this tool will be used in the development of the BSO and in an interface between CPA and BSO (Rumshisky et al., 2006). This tool will allow the core developers of the BSO to finish building the top-level ontology as well as to make more substantial changes to the other portions of the BSO's structure.

4. Acknowledgments

We would like to thank Tim Hickey for advice with respect to the user interface design for such a large and complex system and for his help with the running of the usability testing. Also, the design has been influenced by the indispensable feedback of everyone who has been using the BULB online tools in the past six months. We would like to thank those whose modules allowed us to test the plug-in system in its various stages: Amber Stubbs, Anna Rumshisky and Ben Wellner. Catherine Havasi is funded by a NSF Fellowship.

5. References

Federica Busa, Nicoletta Calzolari, and Alessandro Lenci. 2001. Generative lexicon and the SIMPLE model: Developing semantic resources for NLP. In *The Syntax of Word Meaning*. Cambridge University Press.

O. Christ, B. M. Schulze, A. Hofmann, and E. Konig, editors. 1991. *Corpus Query Processor (CQP)*. Institut fur Maschinelle Sprachverarbeitung, Universitat Stuttgart, Germany.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

Charles Fillmore and Collin Baker. 2000. FrameNet: Frame semantics meets the corpus. In *74th Annual Meeting of the Linguistics Society of America*.

Paul Kingsbury and Martha Palmer. 2002. From Treebank to PropBank. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation*.

A. Meyers, R. Reeves, C. Macleod, R. Szekely, V. Zielinska, B. Young, and R. Grishman. 2004. Annotating noun argument structure for NomBank. In *Proceedings of LREC-2004*, Lisbon, Portugal.

J. Pustejovsky, P. Hanks, and A. Rumshisky. 2004. Automated Induction of Sense in Context. In *COLING 2004, Geneva, Switzerland*, pages 924–931.

James Pustejovsky, Catherine Havasi, Roser Saurí, Patrick Hanks, and Anna Rumshisky. 2005. Towards a generative lexical resource: The Brandeis Semantic Ontology. *Submitted to LREC 2006, Genoa, Italy*.

James Pustejovsky. 1998. *The Generative Lexicon*. MIT Press, Cambridge, MA.

James Pustejovsky. 2001. Type construction and the logic of concepts. In *The Language of Word Meaning*. Cambridge University Press.

Anna Rumshisky, Patrick Hanks, Catherine Havasi, and James Pustejovsky. 2006. Annotating noun argument structure for NomBank. In *Constructing a Corpus-based Ontology using Model Bias*, Melbourne, FL.

Pavel Rychly and Pavel Smrz. 2004. Manatee, Bonito and word sketches for Czech. In *Second International Conference on Corpus Linguisitcs*.

Piek Vossen, editor. 1998. *EuroWordNet: A Multilingual Database with Lexical Semantic Networks.* Kluwer, Dordrecht, Netherlands.