

Unifying Lexical Resources

Dick CROUCH

Palo Alto Research Center
3333 Coyote Hill Rd.
Palo Alto, CA 94304
USA
crouch@parc.com

Tracy Holloway KING

Palo Alto Research Center
3333 Coyote Hill Rd.
Palo Alto, CA 94304
USA
thking@parc.com

Abstract

This paper describes our efforts to create a Unified Lexicon by extracting information from a variety of external resources, namely our XLE syntactic lexicon, WordNet, Cyc, and VerbNet. The UL is built in several steps: first the information is extracted from the resources; then it is merged into lexical entries based on word stem, syntactic subcategorization frame, meaning concept, and WordNet class; finally, patch files are run over the UL to create a cleaner version. The patched version of the UL is used to extract semantics to KR mapping rules, including default rules for where gaps occur in the external resources. This paper focuses on unifying lexical resources for verbs.

1 Introduction

There are a large number of external resources that have been developed to describe different aspects of the syntax, semantics, and abstract knowledge representation of verbs. Since these have been developed at different sites and for different purposes, they contain different types of information in different formats and cover different subsets of English. In order to exploit the information in these resources, it is necessary to merge the information and put it in a uniform format. This paper describes our efforts to build a Unified Lexicon (UL) with lexical entries for verbs based on their syntactic subcategorization in combination with their meaning as described by WordNet, Cyc, and VerbNet.¹

For our purposes, the UL needs to be both machine and human readable. The machine-readability requirement comes from the fact that one of the main goals of these UL entries is to automatically extract rules which map from

¹Other external resources may be incorporated at a later date; these four resources (XLE lexicon, WordNet, Cyc, and VerbNet) were chosen because of their immediate relevance to the sem-kr mapping rules. ULs for other parts of speech are also planned.

syntax to semantics to knowledge representation (Crouch, 2005). A second use of the UL is to determine where there are gaps in the resources and how best to create a series of defaults to fill these gaps. In order to do this, a linguist needs to be able to look through the UL for entries where there is missing information and then find similar entries that can be used to patch this information either with hand-crafted rules or with defaults created from other information, usually information from other entries in the UL.

An sample entry in the UL is shown in fig. 1 for the *HittingAnObject* reading of transitive *hit*, as in *John hit the ball*. The UL entry contains information about WordNet class, Cyc knowledge representation, and VerbNet role mappings, role restrictions, and semantics. Each of these types of information forms a field in the entry, and the content of these fields can be extremely complex (e.g., the VerbNet field in fig. 1). There are also fields for comments, XLE lexicon information other than the subcategorization frame, derivational morphology information, and information from PARC internal resources. Not all fields need to contain information in a UL entry; part of the goal of building the UL is to see where gaps in information arise across the external resources.

The creation and use of the UL involves four steps: the data is extracted from the external resources; the extracted data is merged into the UL entries; the UL entries are corrected with hand-coded and automatically created patch files; mapping rules are extracted from the UL.

2 Extracting the Data

The current UL uses data from the XLE syntactic lexicon, a relatively complete research version of Cyc, VerbNet, and WordNet. We briefly describe these resources and some of the issues that arose when extracting the relevant data from them. In all cases, the data extraction

```

(ul hit v v-SUBJ-OBJ #HittingAnObject
 (wnet ((wn 1172806 (verb contact)) (wn 1198410 (verb contact)) (wn 1359510 (verb contact))))
 (comments ())
 (xle ())
 (cyc (#$and (#$isa ACTION #HittingAnObject) (#$performedBy ACTION SUBJECT)
 (#$objectActedOn ACTION OBJECT)))
 (vnet ( (throw-17_1-1 Basic_Transitive
 ((role SUBJ Agent ((int_control +))) (role OBJ Theme ((concrete +))))
 (sem ((motion (during E1) Theme) (exert_force (during E0) Agent Theme)
 (contact (end E0) Agent Theme) (not (contact (during E1) Agent Theme))
 (cause Agent E1) (meets E0 E1))))))
 (deriv ())
 (parc ()))

```

Figure 1: UL entry for HittingAnObject reading of transitive *hit*

is done automatically to allow us to easily update the UL when new versions of the external resources are released.

The XLE syntactic lexicon is a lexicon associate verb stems (~9,700) with syntactic subcategorization frames (~25,800 stem-frame pairs). It has been developed over the past several years as part of the broad-coverage English LFG grammar for the ParGram project (Butt et al., 2002). Extraction of the data simply comprised extracting each verb stem with its possible subcategorization frames. For example, from the entry in (1), we extract the information that *auction* can be either transitive (*They auctioned the goods*) or transitive with the particle *off* (*They auctioned the goods off/They auctioned off the goods*).

```

(1) auction v
    { @(v-SUBJ-OBJ %stem)
      @(SUBCAT-SOURCE dict)
      |@(v-SUBJ-OBJ_prt %stem off_)
      @(SUBCAT-SOURCE byhand)}.

```

WordNet (Fellbaum, 1998) contains words, in our case only verbs, organized into synonym sets which represent underlying lexical concepts; these synonym sets are linked by relations such as hypernyms (e.g., *auction* is a type of *sell* which is a type of *exchange*, *change*, *interchange* which is a type of *transfer*). WordNet involved basically no direct extraction for the UL. However, WordNet class information is crucially used to determine whether entries from Cyc and VerbNet could be merged (section 3) and the information as to WordNet class(es) is recorded as being potentially useful in other aspects of the system, such as matching across representations. In addition, we anticipate that

WordNet classes will play a crucial role in creating patch files (section 4) to fill in entries where there is not enough information from the other external resources to create useful sem-kr mapping rules.

2.1 Cyc

Cyc is a general knowledge base, including a large ontology of concepts and assertions about these concepts (Lenat, 1995).² Although Cyc contains information about concepts relating to many parts of speech, we initially extracted only the information known to be relevant to verbs. There were three main issues in extracting the Cyc data for inclusion in the UL.

The first concerns lemmatizing the verb forms. Cyc contains not just the base form of the verb, which is what is used in the UL entries, but also many inflected forms (e.g., in addition to listing an entry for transitive *push*, there will be duplicate entries for *pushes*, *pushing*, and *pushed*). To detect these duplicates, we put each verb form through the finite-state inflectional morphology that is used with the XLE English grammar. If this produced a stem with verbal tags that matched an existing verb entry from Cyc, then the form was discarded and only the lemmatized, base one was kept. As discussed in section 4, not all the verbs in Cyc were known to the morphology (e.g., *windsurfed*) and so some inflected entries had to be deleted with patch files.

The second issue involved the encoding of subcategorization frames in Cyc. These frames are labelled as to valency and sometimes phrase-structure type, but not usually with grammat-

²Cyc also contains a reasoning engine which is not used in construction of the UL.

ical functions. For example, the frame #DitransitiveNPCompFrame indicates a verb which takes a subject and two additional NP arguments, such as *I gave him a book*. This must be mapped into grammatical functions as taking a subject, an object, and a secondary/thematic object (SUBJ-OBJ-OBJTH). In some cases, a Cyc frame might map into more than one grammatical function frame. Since there are relatively few frames listed per verb in Cyc, one of the purposes of the UL is to determine what strategies can be used to fill in Cyc-type KR for frames that are not listed. For example, if Cyc only listed the *that*-clause version of a verb and a *wh*-clause version was found in the XLE lexicon and/or VerbNet, could the *that*-clause information from Cyc be reasonably ported to the *wh*-clause one? Strategies for using the UL to fill gaps in external resources like Cyc are the subject of further research; however they must all make use of the patch file mechanism described in section 3.

A final issue with extracting the Cyc data involved the WordNet classes used in Cyc. Cyc associates the relevant WordNet class with a particular meaning of a verb. This information can be used to then associate these meanings with the relevant VerbNet meaning since VerbNet also include WordNet classes (section 2.2). However, two problems arose in doing this. The first was that Cyc uses an older version of WordNet than VerbNet. So, Cyc's WordNet class information had to be converted to the newer WordNet version. A second, more significant problem is that Cyc often uses the WordNet class for the relevant noun instead of verb. Given that Cyc is largely concerned with meaning and hence abstracts away from peculiarities of English syntax, this use of nominal classes for verbs is not unreasonable. However, the WordNet class numbers in these cases could not be used to merge the UL entries. Instead, these verbs had to be looked up in WordNet and then merged based on the retrieved information. The accuracy of the resulting merges is still being assessed, but initial inspection indicates accurate merges.

2.2 VerbNet

VerbNet (Kipper et al., 2000) classifies verbs according to Levin verb classes (Levin, 1993). It includes syntactic subcategorization information, information about thematic roles (e.g., agent, patient), and basic lexical semantics (see

fig. 1). There were three main issues in extracting the VerbNet data for inclusion in the UL.

The first was converting VerbNet subcategorization frames into ones that were compatible with the XLE lexicon. This was difficult because the VerbNet subcategorization information is listed not as grammatical function information but rather as abstractions over the canonical phrase structure tree. For example, the frame corresponding to *present* in *I presented a solution to him*, is represented as in (2) (simplified from the original xml version).

- (2) NP(Agent,[]), verb, NP(Theme,[]), Prep(to,[]), NP(Recipient,[])

From this, we extract the grammatical function specified subcategorization frame V-SUBJ-OBJ-OBL(to). To do this, we determine that the NP before **verb** is a subject, which will be linked to the Agent in the UL representation, and the NP immediately after **verb** is an object, which will be linked to the Theme. The NP following the **Prep** will be an oblique whose prepositional form must be *to* and this oblique will be the Recipient. This extraction becomes extremely involved for verbs which take NP small clauses, particles, expletives, or verbal complements.

The second issue in the VerbNet extraction was ensuring that a verb belonging to a particular VerbNet class inherited all the correct role restrictions from the classes above it. VerbNet classes frequently contain subclasses. Any role restrictions on the class also pertain to the subclass (sometimes nested several deep) and must be extracted accordingly. For example, the *transfer_mesg-37.1* class which applies to sentences such as *Wanda taught French* has a restriction that its Agent is either animate or an organization. The subclass *transfer_mesg-37.1-1* which applies to sentences such as *Wanda taught the students French* and its subclass *transfer_mesg-37.1-1-1* for *Wanda taught the students* both inherit this restriction.

The final issue with VerbNet was that many verb frames have implicit roles. These roles are determined by looking at the semantics provided for the verb. If there is a thematic role mentioned that is preceded by a ?, e.g. ?Topic, then it is implicitly present in the verb frame and may have role restrictions on it. For example, the *transcribe-25.4* class for *The secretary transcribed the speech* has an implicit Destination role which is restricted to being concrete. Note that this role is overt in other frames for

this verb, as in *The secretary transcribed the speech into the record.*

To summarize, extracting the data from external resources into a format that we could then merge into UL entries involved a significant amount of work. Even for someone intimately familiar with all the resources, the conversion would have been non-trivial. Unfortunately, these resources are involved enough that an in-depth understanding of all of them is difficult and so much effort was spent on figuring out what should be extracted, converting it to a uniform format during the extraction, and then doing quality assurance on the results.

3 Merging External Data

Extracting data from a variety of sources and placing it in a moderately uniform format is unfortunately only part of the battle. The data from the different sources needs to be merged. The first stage of merging occurs in data extraction, by virtue of mapping XLE, VerbNet and Cyc verb entries to common subcategorization frames. However, both Cyc and VerbNet make what amount to sense distinctions for individual verbs within a particular sub-categorization frame. The principal task of merging these resources is therefore to identify equivalent Cyc and VerbNet sense distinctions. This is made harder in the case of VerbNet, since following the Levin verb classes it marks semantically significant syntactic alternations rather than alternative senses.

Both VerbNet and Cyc associate verb entries with WordNet synsets. These associations are used to help decide whether to merge Cyc and VerbNet entries for the same verb-subcat frame pairs. Unfortunately, this is not completely straight forward, for a number of reasons. (1) Cyc uses an older release of WordNet than VerbNet. (2) VerbNet uses only verb synsets, while Cyc often associates verbs with relevant nominal synsets. (3) Sense distinctions made by WordNet are often too fine for, and sometimes orthogonal to, ontological distinctions drawn between Cyc.

Part of the merging process attempts to recalculate synsets associated with Cyc entries, as a double check on the WordNet1.6 to 2.1 conversion process. This proceeds by identifying all words, of any part of speech, that map onto a particular Cyc concept, and collecting all the synsets for these words. This forms a very approximate cluster of synsets potentially associ-

ated with a Cyc concept, as opposed to the single synset allocated by Cyc. These clusters give a broader target when trying to match a Cyc entry up with a VerbNet entry. The algorithm is greedy: if a VerbNet entry for a verb with a particular subcat frame has a synset that occurs in the Cyc cluster for the same verb-subcat frame pair, then it is assumed that the two entries should be merged. This sometimes results in multiple matches between a single Cyc entry and VerbNet entries, or vice versa. In such cases, multiple merged entries are produced.

Automated merging of verb senses is error prone. The patch file mechanism described in the next section provides a necessary means for correcting errors.

4 Patching UL Entries

In building the UL, we first extract the data from the relevant sources (XLE lexicon, WordNet, Cyc, and VerbNet) and then merge the information from these resources so that we have one entry for each stem, subcategorization frame, WordNet class, and meaning concept combination. Each such combination forms an id for that UL entry. This initial UL is then modified by patch files. These files can be produced by hand or automatically. The result is a new version of the UL and it is this version that the sem-kr mapping rules are extracted from.

Patch files are a convenient way of keeping a record of changes made to the UL after its initial extraction. It is important that the UL not be hand-edited directly. This is because the external resources from which the UL is constructed are themselves subject to change. We do not want to run the risk of losing hand-made modifications to the UL when rebuilding it to reflect a newer release of one of the external resources. By channeling all modifications to the UL through separate patch files, we can be sure to record any changes made

Patch files can be generated by automatically, semi-automatically or manually. But however they are generated, the format of a patch file is rigidly defined. A patch file consists of an ordered sequence of operations on UL entries, allowing them to be deleted, inserted, merged, or updated. Entries are identified by a key comprising (a) the word stem, (b) the part of speech, (c) the subcategorization frame, (d) the WordNet synset, and (e) a concept index derived from the Cyc knowledge representation of the word. In cases where some of the key infor-

mation is missing (typically the concept index or the synset), null values are used.

Deletion is used to remove entries that are unwanted either because they are incorrect or because they will never be used in the mappings. For example, all of the inflected verb forms from Cyc that were not eliminated in the extraction (e.g., *snowbiking*) are deleted by a patch file. An example of an incorrect reading is that of the intransitive particle verb reading of *nod* for *He nodded off* which is incorrectly listed as #NoddingOnesHead while it should only be listed with the meaning concept associated with falling asleep.

Insertion occurs when an entirely new entry is needed. Often, updating is used instead of insertion because existing underspecified UL entries, e.g. ones only with XLE lexicon and WordNet information such as *abbreviate* and *abdicate*, can be updated with the relevant additional information.

Merge merges two or more existing UL entries into a single new entry. Merges may be necessary where the WordNet classes did not align perfectly and yet the intended meanings of the two entries are identical. Often, an update to a UL entry will result in a new entry which can then be merged into an existing one. For example, if an incorrect subcategorization frame has been extracted from Cyc, this frame can be updated to the correct one and then the Cyc entry can be merged with an existing VerbNet one. This is done for many verbs taking prepositions since the encodings in Cyc and VerbNet were sometimes ambiguous between verbs taking obliques and those taking particles. In such cases, both were hypothesized in the original extraction and then updated and merged with a patch file based on the correct analysis.

Updating replaces a specified field in the UL with a new one.³ Three operations are possible: adding, removing, and replacing. Each of these operates on a specified field in the UL. The fields include the word itself, the subcategorization frame, each of the types of extracted information (e.g., VerbNet), and a comment field. Adding creates a value for a field where there was none before. This can be used to insert comments into the UL entry. For example, many VerbNet entries have oblique arguments

³There is also an update and copy command that copies the entry and then only updates the copy, leaving the original entry as well. This is often used to split entries and then merge them with several other entries.

that the XLE grammar analyzes as adjuncts. For example, the XLE lexicon has a transitive use of *punch* but not one which takes an object and an *on* oblique (e.g., *He punch him on the arm*). For these verbs, a comment is inserted stating that the oblique is an adjunct and this comment allows the extracted sem-kr mapping rules to look for the appropriate adjunct grammatical function instead of an oblique one. Removing deletes the information in a given field. For example, if the VerbNet information for a given verb was incorrect but the Cyc and XLE information was correct, the UL entry could be updated by removing the VerbNet field. Finally, replacing removes the existing value for a field and replaces it with a new one. This is used to turn the Cyc multiword verbs into their single word equivalents. For example, the word *breathe in* is replaced by *breathe* and simultaneously its intransitive subcategorization frame is replaced by the intransitive frame with an *in* particle. This new entry can then be merged with the existing UL entry for that reading of the verb.

To summarize, a system of patch files is available to modify the UL from its initial state in which only information extracted from the external resources is used. Patch files can delete, insert, and merge entries, as well as modify any field in the entry. Since the rules in the patch file are ordered, entries are often modified and then merged to create single, accurate UL entries with information unified from all of the external resources.

5 Results and Conclusions

This paper describes our efforts to create a Unified Lexicon by extracting information from a variety of external resources, namely the XLE syntactic lexicon, WordNet, Cyc, and VerbNet. The UL is built in several steps: first the information is extracted from the resources; then it is merged into lexical entries based on verb stem, syntactic subcategorization frame, meaning concept, and WordNet class; finally, patch files are run over the UL to create a cleaner version. The patched version of the UL is then used to extract sem-kr mapping rules, including default rules for where gaps occur in the external resources.

The current UL contains 45,704 entries for 9,835 verb lemmata. 22,208 have no VerbNet information. 42,160 have no Cyc information. Of these, 22,122 have neither VerbNet nor Cyc

information (e.g., *adapt*); that is, they effectively only contain the information from the XLE syntactic lexicon and WordNet. 17,991 have syntactic frames which came from VerbNet and were not in the XLE lexicon; the majority of these are frames with multiple oblique PP arguments (e.g., *The witch turned him from a prince into a frog*) and various types of resultatives (e.g., *Linda taped the box shut*) and middles (e.g., *Labels tape easily to that kind of cover*).

There is still much work to be done to fully exploit the UL in our syntax to semantics to KR mapping system. The next task is to extract mapping rules from the UL and incorporate them into the sem-kr mapping system. Then patch files need to be created to systematically fill in some of the gaps in the UL. Since there are many entries with VerbNet information but no Cyc information, we hope to use the VerbNet information to make informed guesses as to the Cyc meaning of the verb. In addition, WordNet classes may be used to determine the closest synonym for a given verb and the entry for that synonym could then be used to augment the UL entry for that verb.

Longer term work includes the incorporation of other external resources into the UL (e.g., derivational morphology, ComLex, FrameNet). In addition, ULs are being created for other parts of speech, including nouns and adjectives. The immediate need for these in our system is less pressing than for the verb UL described here because the external resources, in particular Cyc, can be used directly as a temporary measure.

6 Acknowledgements

This work was supported in part by the Advanced Research and Development Activity (ARDA)'s Advanced Question Answering for Intelligence (AQUAINT) Program.

References

- Miriam Butt, Helge Dyvik, Tracy Holloway King, Hiroshi Masuichi, and Christian Rohrer. 2002. The Parallel Grammar project. In *Proceedings of COLING2002, Workshop on Grammar Engineering and Evaluation*.
- Dick Crouch. 2005. Packed rewriting for mapping semantics to KR. In *Proceedings of the 6th International Workshop on Computational Semantics*.
- Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. The MIT Press.
- Karin Kipper, Hoa Trang Dang, and Martha Palmer. 2000. Class-based construction of a verb lexicon. In *AAAI-2000 17th National Conference on Artificial Intelligence*.
- Doug Lenat. 1995. Cyc: a large-scale investment in knowledge infrastructure. In *CACM 38, No. 11*.
- Beth Levin. 1993. *English Verb Classes and Alternations*. Chicago University Press.