The Cross-Breeding of Dictionaries

Adam Meyers, Ruth Reeves, Catherine Macleod, Rachel Szekely,
Veronika Zielinska, Brian Young

New York University
719 Broadway, 7th Floor
New York, New York 10003
U.S.A.
meyers/reevesr/macleod/szekely/zielinsk/byoung@cs.nyu.edu

Abstract

Especially for English, the number of hand-coded electronic resources available to the Natural Language Processing Community keeps growing: annotated corpora, treebanks, lexicons, wordnets, etc. Unfortunately, initial funding for such projects is much easier to obtain than the additional funding needed to enlarge or improve upon such resources. Thus once one proves the usefulness of a resource, it is difficult to make that resource reach its full potential. We discuss techniques for combining dictionary resources and producing others by semi-automatic means. The resources we created using these techniques have become an integral part of our work on NomBank, a project with the goal of annotating noun arguments in the Penn Treebank II corpus (PTB).

1. Introduction

Especially for English, the number of hand-coded electronic resources available to the Natural Language Processing Community keeps growing: annotated corpora, treebanks, lexicons, wordnets, etc. Over the last decade, virtually every professional conference has had at least one talk describing a new -bank, a new -net or a new -lex. Unfortunately, initial funding for such projects is much easier to obtain than the additional funding needed to enlarge or improve upon such resources. Thus once one proves the usefulness of a resource, it is difficult to make that resource reach its full potential. This paper discusses techniques for combining dictionary resources and producing others by semi-automatic means. This makes it possible to enrich existing resources while building new ones efficiently. This paper describes several resources that we created and/or enriched by combining automatic and manual approaches. These resources have become an integral part of our work on NomBank, a project with the goal of annotating noun arguments in the Penn Treebank II corpus (PTB). NomBank is part of the larger effort to add logical and semantic levels of annotation to the Penn Treebank. The first part of that effort to be completed was PropBank (Kingsbury et al., 2002; Kingsbury and Palmer, 2002).

2. Resources to Start With

We began with the following hand built resources:

- COMLEX Syntax (Macleod et al., 1998a) – a syntactic dictionary of nouns, verbs, adjectives and adverbs.
- NOMLEX (Macleod et al., 1998b) – a dictionary listing 1000 nominalizations, their related verbs, and correspondences between the verbal arguments and syntactic positions within the noun phrase.
- PropBank's Frame Dictionary (Kingsbury and Palmer, 2002) — Lexical entries providing verbal argument structure for The University of Pennsylvania's PropBank Project
- The Verb Index from (Levin, 1993) – a downloadable index of the verb classes described in the cited work.
- CATVAR (Habash and Dorr, 2003) – A dictionary pairing up lexical items related by derivational morphology. The creators of CATVAR used both previous resources and automatic procedures.

3. Sketchy Dictionaries and Word Lists

We also created several sketchy dictionaries initially by automatic means, but then hand edited them, deleting items and classifying others with simple labels. These sketchy wordlists were used as the basis of more structured entries. They included lists of:

1. Potential nominalizations and corresponding verbs from COMLEX
2. Morphologically related adjective/adverb pairs from COMLEX
3. Morphologically related adjective/noun pairs from COMLEX
4. Verbs that take atypical subjects

The nominalization/verb list (1) and adjective/adverb list (2) were created by checking for noun/verb and adjective/adverb pairs with large shared prefix strings, e.g., the pair destruction/destroy share the prefix *destr-*. Same strings were collected as well as prefixes with specified pairs of suffixes. For example, anesthetist/anesthetize share the prefix *anesthet* and match the suffix pair -ize/-ist and slow/slowly share the suffix pair NULL/-ly. Other morphological rules were also applied so that some near pairings would be allowed, e.g., a final "i" was assumed to match a final "y" in a pair of prefixes. This technique overgenerates somewhat producing odd pairs like *secretary/secrete* and we edited the results by hand to compensate for this. For the adjective/noun list (3), we used a different method. We classified a subset of the nouns in the PTB by hand and then extended the pairs by analogy: from each pairing

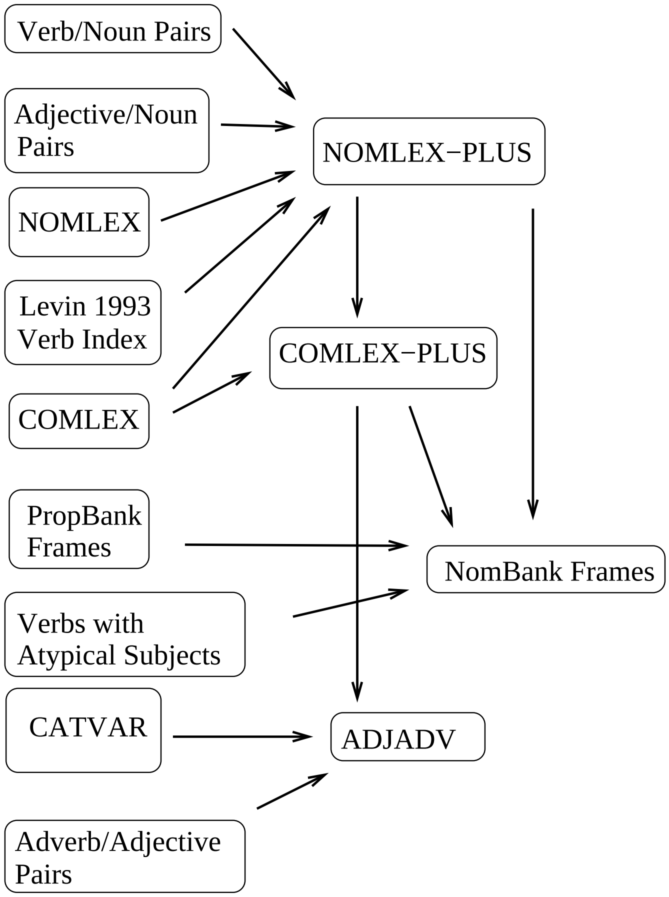

Figure 1: The Relation Between Dictionaries

of an adjective and its related nominalization, we extracted the pair of endings that mark the differences between these two words; then we used all such pairings to derive additional adjective/nominalization pairs. For example, the pair of endings in *-ability/-able* can be extracted from the pair *durability/durable* and then used to identify the pair *availability/available*. The nominalization list was also extended by this same method.

We created a list of verbs that take atypical (e.g., themes/patients) subjects (4) in order to get accurate role assignment. As we already had a list of alternating verbs (the Levin classes), we focused on intransitive verbs. We started with the set of verbs from COMLEX Syntax that meet the following criteria: (a) the verb can occur with no complement (INTRANS); (b) the verb cannot take a simple NP complement (NP); and (c) there is some nominalization of that verb in our database. We then edited this list, keeping only verbs that took atypical subjects. This last limitation was artifact of our task – we were only concerned with arguments of nominalizations.

4. The New Resources

We used the above word lists and previously constructed resources to create the new resources described in this section. Figure 1 shows the relationships among the dictionaries and word lists above and the new resources described below. Arrows indicate a “derived from” relation.

4.1. NOMLEX-PLUS

NOMLEX-PLUS is a 7050 entry extension of NOMLEX (it includes the original 1000 entries of NOMLEX). NOMLEX-PLUS has 4900 entries for nominalizations of verbs, 550 entries for nominalizations of adjectives and 1600 entries that fall into 16 classes for argument-taking nouns including PARTITIVE nouns (*a VARIETY of books*), RELATIONAL nouns (*PRESIDENT of the company*, *Mary’s FATHER*), ATTRIBUTE nouns (*the VOLUME of the sphere*), among others.

Beginning with our semi-automatically classified nominalizations of adjectives and verbs, we label the remaining common nouns in the PTB as either some type of nominalization, one of 16 other classes of argument-taking nouns (relational noun, partitive, etc.) or as nouns that do not take arguments, in which case they are excluded from the dictionary. Default noun argument to verb argument mappings were then used to create NOMLEX-style entries to record how syntactic positions within the NP are filled by particular argument types. Additional information was added to mark verb alternations (using the Levin verb index) and similar information was added by hand.

These 5450 nominalization entries include the original 1000 NOMLEX entries. The main differences between the additional 4450 NOMLEX-PLUS entries and the entries from its predecessor are:

1. The NOMLEX entries were created by hand, whereas the NOMLEX-PLUS entries were created as described above. This means that the mappings in the NOMLEX entries tend to be more accurate. In contrast, the NOMLEX-PLUS entries reflect a set of defaults associated with the complement classes in the COMLEX entry of the related verb and various other factors, e.g., some of the classes in (Levin, 1993) are taken into account. For example, by default, simple transitive (NOM-NP) complements allows: (a) both the verbal subject and object to occur in possessive or prenominal modifier position; (b) allows the verbal subject to occur as the object of the preposition *by*; and (c) allows the verbal object to occur as the object of the preposition *of*.

2. NOMLEX only lists nouns that are related morphologically to verbs, whereas NOMLEX-PLUS includes nouns that take arguments like nominalizations, but are not morphologically related to any verbs. These “cousins” of nominalizations were manually associated with verbs with similar argument taking properties. For example, the entries for *ado* and *anniversary* were based on the entries for *fuss* and *commemorate*.

3. NOMLEX-PLUS includes nominalizations of adjectives. As with the nominalization of verb entries, argument assignment was created by a system of defaults based on the COMLEX Syntax entry for the adjective.

Similarly, a NOMLEX-like entry was provided for each of the 16 noun classes. These entries were based on nominalizations that belonged to a particular class. For example, the entries for partitives were based on the entries of the

```
(NOM
   :ORTH            "abduction"
   :VERB            "abduct"
   :NOM-TYPE        ((VERB-NOM))
   :VERB-SUBC
     ((NOM-NP       :SUBJECT
       ((DET-POSS)(N-N-MOD)(PP :PVAL (”by”)))
                    :OBJECT
       ((DET-POSS)(N-N-MOD)(PP :PVAL (”of”)))))))
```

Figure 2: Simplified NOMLEX-PLUS entry for *abduction*

nominalizations *variety* and *cascade* and entries for RELATIONAL nouns were based on entries for *teacher*, *leader* and *director*. This means that our entries for relational nouns are like entries for subject nominalizations. Furthermore, it means that to the extent possible, if a nominalization belongs to one of the 16 classes, the nominalization entry and the noun class entry will make mostly the same predictions about argument structure. This redundancy is desirable because it means that the dictionary will handle nouns with similar argument in similar ways, even if one noun is a nominalization, e.g., *variety* and another is a partitive noun, e.g., *multitude*. In a sense, the classes may be thought of as standardized sets of "cousins" of nominalizations.

A simplified NOMLEX-PLUS entry is provided for *abduction* in figure 2. This is based on the fact that the verb *abduct* takes an NP complement in COMLEX Syntax and our defaults for that complement class.

4.2. ADJADV

ADJADV is a dictionary defining adverbial uses of adjectives. For example, *possible* in *possible abduction* has a meaning similar to the epistemic adverb *possibly*. Similarly, the word *fine* has the same evaluative meaning regardless of whether it is used adjectivally (*fine behavior*) or adverbially *He behaved fine*. We began with all the adjectives in the Penn Treebank Corpus (regularized for -er and -est inflection) and we pulled out every adjective that was associated with an adverb either by the list we created at NYU (this accounted for the -ly inflection and adjective/adverb pairs that had the same orthography, e.g., *fine*) or was associated with some adverb by CATVAR. In some cases, manual inspection showed that a different adverb should be associated with the adjective than predicted by these word lists. For example, we recognize the adjective *awesome* in *his awesome performance* has a similar meaning to the adverb *amazingly* in *He performed amazingly*, but has little in common with the adverb *awful* as predicted by the more automatic means. Thus we derive an adverb-like entry for *awesome* based on COMLEX's entry for *amazingly*. A sample ADJADV entry is provided as figure 3. Possible values for :FEATURES are a subset of the ones for adverbs in COMLEX Syntax.

4.3. COMLEX-PLUS

COMLEX Syntax dictionary has over 100 complement classes for verbs, but much fewer for nouns. In particular, it lacks PP complements for nouns. Fortunately, many of the nouns that take PP and other complements are found in NOMLEX-PLUS. We can therefore apply some simple rules for adding PP complements to the nouns in COMLEX. Similarly, we can add any missing clausal arguments for nouns. Our procedures used the postnominal noun argument positions referenced in each NOMLEX-PLUS entry to augment the corresponding noun in COMLEX. For example, the NOMLEX-PLUS entry for *abduction* in figure 2 would cause our procedures to add PP complements headed by *of* and *by* to form the COMLEX-PLUS entry in figure 4.

```
(ADJADV
   :ORTH "possible"
   :ADV "possibly"
   :FEATURES ((META-ADV :EPISTEMIC T)))
```

Figure 3: Sample ADJADV entry for *possible*

```
(NOUN    :ORTH "abduction"
         :SUBC ((PP :PVAL ("of" "by"))))
```

Figure 4: COMLEX-PLUS entry for *abduction*

4.4. NomBank Frame Dictionary

NOMLEX-PLUS, the University of Pennsylvania's verb frames and other information such as our list of verbs with atypical subjects were used to automatically generate lexical entries for all argument-taking nouns in the PTB. These entries provide an inventory of the role labels (ARG0, ARG1, ...) which occur for particular nouns (in simple cases these correspond to subject, object, indirect object, etc). These are now being used as our initial lexical entries for NomBank, although annotators modify them as needed. Figure 5 is a simplified NomBank lexical entry for *abduction*.[1] In this case, the frame is a mirror of the verb frame for *abduct*.

5. Using These Resources: Present and Future

The NomBank Frame Dictionary is a necessary part of the NomBank annotation project. By automatically creating initial versions of these lexical entries, we are greatly

```
(PBNOUN
   :ORTH "abduction"
   :ROLE-SETS
    ((ROLE-SET1
      :ID "abduction.01"
      :SOURCE "verb-abduct.01"
      :NAME "TO CARRY SOMEONE OFF BY FORCE"
      :ROLES
         ((ROLE :DESCR "AGENT" :N "0")
          (ROLE :DESCR "PATIENT" :N "1")))))
```

Figure 5: Sample NomBank Lexical Entry for *abduction*

[1] This figure uses lisp-like format for compatibility with the other lexical entries presented here. However, there is an equivalent XML format for use with NomBank.

increasing the speed at which these entries can be created. In fact, we err on the side of overgenerating choices of rolesets rather than undergenerating. The annotators can then delete some of the choices when they see what actually occurs (it is much easier to delete text than to create new text). As discussed above, many of the previous resources were involved in creating this resource. Sometimes, annotator feedback is used to improve the mapping procedure. In addition, annotators can help edit the resulting resources. For example, as a side-effect of the NomBank project, annotators have contributed to improving both accuracy and coverage of ADJADV and NOMLEX-PLUS. This is in addition to their work on the NOMBANK frame dictionary, which is an integral part of the NOMBANK project.

In some related work, we intend to use all of these resources as part of an effort to automatically produce predicate argument structure from Penn Treebank II format text (either the Penn Treebank itself or Penn Treebank-based parser output). This research will surface as both: (a) part of GLARF (Meyers et al., 2001a; Meyers et al., 2001b; Meyers et al., 2002), a formalism and set of mapping procedures for producing a typed feature structure representation of predicate argument structure; and (b) automatic NomBank annotation. For example, given the NP, *her possible abduction* to derive the proposition:

REL = abduction, ARG1 = her, ARGM-MNR = possible

which could be paraphrased as “Possibly, somebody abductor her”. The above dictionary entries provide sufficient information to automatically identify *her* as the object or ARG1 of *abduction*. The DET-POSS or possessive position is an option for both object and subject position in figure 2. Furthermore, both slots can be filled by a human (*her*) on selectional grounds as indicated by the “AGENT” and “PERSON KIDNAPPED” :DESCR features in the ARG0/ARG1 slots in figure 5. Nevertheless, barring selection restrictions, this sort of ambiguity is usually resolved in favor of the object position, barring other considerations, e.g., in the original NOMLEX, additional features can be stated that override this tendency. Note that other information can force a subject reading, e.g., in *her abduction of the puppy*, the *of* phrase can only be interpreted as an object. Thus only the subject role can be reasonably assigned to *her*.

We will use GLARF output to automatically produce NomBank annotation. We will use this automatically produced annotation as both a preprocessor for human annotation and as a tool for finding errors in human annotation. The human annotator will have a chance to survey the quality of the automatically produced annotation before using it. If the annotator decides that the quality is high enough, he/she will edit the automatically produced annotation rather than starting from scratch. If the automatic annotation contains too many errors, we may compare it with the human annotator output as an aid for error detection.

6. Summary

We have outlined a technique for using previously produced dictionary resources to update each other and to produce new resources, allowing for some human intervention. We have briefly described how we have applied this technique in the context of our work at the NomBank project at New York University.

Acknowledgments

Nombank is supported under Grant N66001-001-1-8917 from the Space and Naval Warfare Systems Center San Diego. This paper does not necessarily reflect the position or the policy of the U.S. Government.

7. References

Habash, N. and B. Dorr, 2003. CatVar: A Database of Categorial Variations for English. In *Proceedings of the MT Summit*. New Orleans.

Kingsbury, P. and M. Palmer, 2002. From treebank to propbank. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC-2002)*. Las Palmas, Spain.

Kingsbury, P., M. Palmer, and Mitch Marcus, 2002. Adding semantic annotation to the penn treebank. In *Proceedings of the Human Language Technology Conference*. San Diego, California.

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

Macleod, C., R. Grishman, and A. Meyers, 1998a. COMLEX Syntax. *Computers and the Humanities*, 31(6):459–481.

Macleod, C., R. Grishman, A. Meyers, L. Barrett, and R. Reeves, 1998b. Nomlex: A lexicon of nominalizations. In *Proceedings of Euralex98*.

Meyers, A., R. Grishman, and M. Kosaka, 2002. Formal Mechanisms for Capturing Regularizations. In *Proceedings of LREC-2002*. Las Palmas, Spain.

Meyers, A., R. Grishman, M. Kosaka, and S. Zhao, 2001a. Covering Treebanks with GLARF. In *ACL/EACL Workshop on Sharing Tools and Resources for Research and Education*.

Meyers, A., M. Kosaka, S. Sekine, R. Grishman, and S. Zhao, 2001b. Parsing and GLARFing. In *Proceedings of RANLP-2001*. Tzigov Chark, Bulgaria.