Using NOMLEX to Produce Nominalization Patterns for Information Extraction

Adam Meyers, Catherine Macleod, Roman Yangarber,
Ralph Grishman, Leslie Barrett, Ruth Reeves
New York University
715 Broadway, 7th Floor, NY, NY 10003, USA
meyers@cs.nyu.edu

Abstract

This paper describes how NOMLEX, a dictionary of nominalizations, can be used in Information Extraction (IE). This paper details a procedure which maps syntactic and semantic information designed for writing an IE pattern for an active clause (*IBM appointed Alice Smith as vice president*) into a set of patterns for nominalizations (e.g., *IBM's appointment of Alice Smith as vice president* and *Alice Smith's appointment as vice president*).

1 Introduction

Although, nominalizations[1] are very common in written text, the computational linguistics literature provides few systematic accounts of how to deal with phrases containing these words. This paper focuses on this problem in the context of Information Extraction (IE).[2] Many extraction systems use either parsing combined with some form of syntactic regularization, or a meta-rule mechanism to automatically match variants of clausal syntactic structures (active main clause, passive, relative clause etc.), e.g., FASTUS (Appelt et al., 1995) and the Proteus Extraction System (Grishman, 1995). However, this mechanism does not extend to nominalization patterns, which must be coded separately from the clausal patterns. NOMLEX, a dictionary of nominalizations currently under development at NYU, (Macleod et al., 1997) provides a way to handle nominalizations more automatically, and with greater coverage. NOMLEX includes information about mappings between verbs and nominalizations that will help generalize information from verbal patterns to create nominalization patterns. This paper describes the structure of the dictionary and a procedure for creating nominalization patterns. This procedure takes into account lexical information about nominalizations which is encoded in NOMLEX.

The Proteus Extraction System starts with a semantic pattern for an active clause:

np(C-company) vg(appoint) np(C-person) "as" np(C-position)

which matches a clause beginning with a noun phrase headed by a noun of type COMPANY, followed by a verb group (verb plus auxilliaries) headed by *appoint*, a noun phrase headed by a noun of type PERSON, the literal *as*, and a noun phrase headed by a noun of type *position*, e.g., *IBM appointed Alice Smith as vice president.* Proteus applies meta-rules to this pattern to produce new patterns for other clausal types, e.g., a passive clause:

np(C-person) vg-pass(appoint) "as"
np(C-position) "by" np(C-company)

(vg-pass is a passive verb group). This new pattern would match *Alice Smith was appointed as vice president by IBM.* When a pattern matches input text, the pieces of the text corresponding to the constituents of the pattern are used to build a semantic representation of the text.

To avoid the need for having users code such patterns, we have developed the Proteus Extraction Tool (PET) (Yangarber and Grishman, 1997). PET allows the user to input an example sentence and specify the mappings from syntactic to semantic form. The system then produces

[1] Nominalizations are nouns which are related to words of another part of speech, most commonly verbs. In this paper, only verbal nominalizations will be discussed.

[2] The Message Understanding Conference Scenario Template Task (MUC, 1995), (MUC, 1998) is our model for the kind of information that we are attempting to extract (who does what to whom, and when and where).

generalized patterns to perform these mappings. This paper shows how PET can use NOMLEX to create nominalization patterns as well. For example, given the sentence *IBM appointed Alice Smith as vice president*, human input and dictionary entries identify IBM as the employer, Alice Smith as the employee, and vice president as the position. The meta-rules add a slot for temporal PPs which state the date (e.g., *on June 1, 1998*). PET creates patterns to fill the semantic slots (employer, employee, position) from the gramatical roles (subject, object, NP following *as*, etc.) in the sentence. PET generates patterns to cover passive sentences, active sentences and relative clauses. Enhanced with NOMLEX, PET can also cover examples like *Alice Smith's appointment as vice president*; *IBM's June 1, 1998 appointment of Alice Smith*; and *the June 1, 1998 appointment of Alice Smith by IBM.* The correspondence between nominal and verbal positions is determined by explicit information in the NOMLEX dictionary entry and by general linguistic constraints.

2 Considerations for Choosing a Dictionary Encoding

The primary information in a NOMLEX entry is a description of a nominalization's argument structure. This information can be quite complex. There are several potential argument positions, including both pre-nominal (*the bomb explosion*, *the bomb's explosion*) and post-nominal (*the explosion of the bomb*), and a given verbal argument may appear in one of several positions.[3] In general, individual arguments may be omitted, although there are some co-occurrence constraints, which we shall consider below. Furthermore, whether or not one position is filled may affect the interpretation of other positions; thus, in *Rome's destruction*, *Rome* is the object, whereas in *Rome's destruction of Carthage*, *Rome* is the subject.

In seeking an appropriate representation for this information, one can compare the situation with that of English verbal complements, which have been extensively studied and recorded. In English, verbal complements are relatively fixed in composition and order. As a result, the common practice (adopted, for example, in OALD (Hornby, 1980), LDOCE (Proctor, 1978), and COMLEX Syntax (Macleod et al., forthcoming) is to enumerate and name the possible subcategorizations, where in general each subcategorization represents a fixed sequence of syntactic elements.[4] For example, in COMLEX (Wolff et al., 1994), NP-PP consists of a Noun Phrase followed by a Prepositional Phrase as in *put the milk in the refrigerator*, where **put the milk*, **put in the refrigerator* and **put in the refrigerator the milk* are not acceptable. Such an approach would be unwieldy for nominalizations, where an argument may appear in several positions and may also be omitted. As a result, even a simple nominalization may entail a large number of subcategorization frames. If these were listed explicitly, the entry would be difficult to create and to read.

On the other hand, a representation which separately listed all the complement structures which could occur with a nominalization, assuming they could freely co-occur, would fail to capture many crucial constraints between complements. For example, the nominalization *confirmation* has both THAT-S (*His confirmation THAT HE WOULD GO*) and WH-S complements (*His confirmation of WHETHER HE WOULD GO.*). However, these complements cannot co-occur (**His confirmation THAT HE WOULD GO of WHETHER HE WOULD GO.*). Also in the case where the associated verb has an NP-AS-NP complement (*She treated them as inferiors*) and no AS-NP complement (*She emerged as their main competitor*) the nominalization cannot have a bare AS-NP. Thus we have *The consideration of HIM AS A CANDIDATE* and *HIS consideration AS A CANDIDATE* but not **The consideration AS A CANDIDATE.*

Guided by these considerations, we chose an approach in which we first determine which COMLEX verbal complements can correspond

[3] These positions may be filled by non-arguments as well. For example, the prenominal positions may be filled by temporal NPs (NTIME1 and NTIME2 in COMLEX) like *Yesterday* (but not *John*), e.g., *Yesterday's appointment of the Prime Minister* and *The June 1, 1987 appointment of the Prime Minister.* These positions correspond to temporal modifiers of clauses, e.g., *X appointed the Prime Minister Yesterday/June 1, 1987.* For further discussion, see Section 5.

[4] In COMLEX Syntax, some symbols designate sets of alternative complement structures, e.g., the ditransitive alternation.

to phrases containing nominalizations and then we specify how these complements can be mapped to arguments of the nominalizations. The resulting COMLEX-based encoding does not permit incompatible complement phrases to co-occur, e.g., *confirmation* would not simultaneously take both THAT-S and WH-S complements. Optionality, obligatoriness and alternative positions of phrases is stated in a simple notation, e.g., it can be stated in the entry for *consideration* that the verbal object for the NP-AS-NP complement of *consider* maps to either the DET-POSS position (*HIS consideration as a candidate*) or the PP-OF position (*The consideration OF HIM as a candidate*) and that this object is obligatory for mappings of NP-AS-NP, i.e., if the object is not present in the phrase containing *consideration*, then the phrase cannot be mapped to the NP-AS-NP complement, although other complements are possible.

Our representation also accounts for the difference in behavior of the core arguments (the subject, the object, and the indirect object) and the other arguments, which we shall refer to as oblique complements. The core arguments, as we have noted, can appear in several positions in the nominalization, and may be independently omitted or included. The oblique complements of the verb, on the other hand, generally translate directly into nominalization complements, either unchanged or occasionally with the addition of a preposition or a "that" complementizer.

3 What is a NOMLEX Entry

NOMLEX entries are organized as typed feature structures and written in a Lisp-like notation (Figures 1 and 2). Each entry lists the nominalization (:ORTH) and the associated verb (:VERB). The :NOM-TYPE feature specifies the type of nominalization: VERB-NOM for nominalizations which represent the action (*destruction*) or state (*knowledge*) of the associated verb; VERB-PART for nominalizations which incorporate a verbal particle (*takeover*); SUBJECT for nominalizations which represent the subject of the verb (*teacher*), and OBJECT for nominalizations which represent the object of the verb (*appointee*). The :NOUN keyword includes information about whether the word has non-nominalization noun senses (and may include some frequency information). For example, *appointment* has a sense which means something like "date to do something". We are only interested in the nominalization sense in this paper.[5]

The heart of the entry is a list of verb subcategorizations, :VERB-SUBC, taken from COMLEX Syntax. The name for each subcategorization is prefixed by NOM- (such as NOM-NP or NOM-NP-AS-NP) and, for subcategorizations involving prepositions, :PVAL specifies those prepositions. The COMLEX complements in these lexical entries include:

- NP, a noun phrase complement, e.g., *IBM appointed Mary*
- NP-PP, a complement consisting of a noun phrase and prepositional phrase, e.g., *IBM appointed Mary for the vice presidency*
- NP-TO-INF-OC, a complement consisting of a noun phrase object and an infinitive clause, where the subject of the infinitive corresponds to the object of the main clause, e.g., *IBM appointed Mary to do the job*
- NP-AS-NP, a complement consisting of a noun phrase object, the word "as" and a second noun phrase, e.g., *IBM appointed Mary as vice president*

For each verb complement, the entry lists the associated nominalized structure, if different from the verbal complement. The entry also lists the positions in which the object (:OBJECT) may appear. For *appointment*, these positions include the following for most complements:

- DET-POSS, a possessive determiner, e.g., *Alice Smith's appointment as vice president*
- N-N-MOD, a nominal prenominal modifier, e.g., *the Alice Smith appointment for vice president*
- PP-OF, object of the preposition *of*, e.g., *the appointment of Alice Smith*

[5] When two or more argument positions are filled, the semantic classes of the arguments in the examples limit our patterns to the nominalization senses. However, patterns in which only one argument position is filled may match phrases that are ambiguous, e.g., *Alice's appointment* can refer to either a dental appointment or an appointment to the vice presidency. These cases are handled by other modules of Proteus, such as inference rules or reference resolution.

```
(NOM  :ORTH "appointment"
      :VERB "appoint"
      :PLURAL "appointments"
      :NOUN (exists)
      :NOM-TYPE (VERB-NOM)
      :VERB-SUBJ ((N-N-MOD) (DET-POSS))
      :SUBJ-ATTRIBUTE (COMMUNICATOR)
      :OBJ-ATTRIBUTE (NHUMAN)
      :VERB-SUBC
        ((NOM-NP :OBJECT ((DET-POSS) (N-N-MOD) (PP-OF))
                  :REQUIRED ((OBJECT)))
         (NOM-NP-PP :OBJECT ((DET-POSS) (N-N-MOD) (PP-OF))
                  :PVAL ("for" "to") :REQUIRED ((OBJECT)))
         (NOM-NP-TO-INF-OC :OBJECT ((DET-POSS) (PP-OF))
                  :REQUIRED ((OBJECT)))
         (NOM-NP-AS-NP :OBJECT ((DET-POSS) (PP-OF))
                  :REQUIRED ((OBJECT)))))
```

Figure 1: NOMLEX entry for *appointment*

```
(NOM  :ORTH "appointee"
      :VERB "appoint"
      :PLURAL "appointees"
      :NOM-TYPE (OBJECT)
      :VERB-SUBJ ((PP-OF) (NOT-PP-BY) (N-N-MOD) (DET-POSS))
      :SUBJ-ATTRIBUTE (COMMUNICATOR)
      :OBJ-ATTRIBUTE (NHUMAN)
      :VERB-SUBC
        ((NOM-NP)
         (NOM-NP-PP :PVAL ("for" "to"))
         (NOM-NP-AS-NP)))
```

Figure 2: NOMLEX entry for *appointee*

However, NOM-NP-AS-NP does not allow the N-N-MOD position (* *the Alice Smith appointment as vice president*). The :OBJECT is not indicated for the :VERB-SUBC of *appointee* because the nominalization itself corresponds to the verbal object (it is :NOM-TYPE ((OBJECT))).

Because a subject argument can appear with any verbal complement, we include, at the top level, a list of positions for the subject (:VERB-SUBJ). This list can be further restricted for a particular complement by including a :SUBJECT feature under that complement in the NOMLEX entry. As a default, it is assumed that subjects always can map to prepositional *by* phrases. Exceptions are marked with NOT-PP-BY, as in the entry for *appointee* (**the appointee by IBM (for vice president)*).

Typically, a nominalization will list multiple positions for each core argument. This doesn't mean, however, that all combinations of the positions are possible. Several constraints limit the possible role assignments; some of these constraints are general, and some are based on particular lists in each entry:

- The uniqueness constraint says that any verbal role may only be filled once. For example, in *Leslie's appointment of Alice*, the PP-OF position filled by *Alice* must map into the object role. As a result, *Leslie* cannot fill the object role and therefore must fill the subject role.[6]

[6] This constraint is based on the stratal uniqueness theorem of (Johnson and Postal, 1980) and related work in Relational Grammar, which is assumed to be a constraint across all languages.

- The ordering constraint says that, if there are multiple pre-nominal modifiers, they must appear in the order subject, object, indirect object, and oblique; thus, for example, *John Smith's school board appointment* cannot mean that the school board appointed John Smith to some position. [7]
- Some entries contain :SUBJ-ATTRIBUTE and :OBJ-ATTRIBUTE features, selectional constraints which are useful in selecting role assignments. In Figures 1 and 2, the attributes COMMUNICATOR (organization, person, or other entity capable of communicating) and NHUMAN (a human) are used.
- Obligatoriness constraints are assumed for the mappings associated with each complement in each entry. As a default, it is assumed that only subject and object are optional. Therefore, *Mary Smith's appointment* would be associated with the NOM-NP complement, but not the NOM-NP-AS-NP complement. Furthermore, objects are obligatory for a particular complement *NP-X* for a particular nominalization *N*, if *N* takes both *NP-X* and *X* as complements, where *NP-X* includes all the phrases in *X* plus an object (e.g., NOM-NP vs. NOM-INTRANS, NOM-NP-PP vs. NOM-PP, NOM-NP-THAT-S vs. NOM-THAT-S, etc.). These defaults can be overridden in the dictionary with attributes on specific complements specifying which roles are :OPTIONAL or :REQUIRED. For example, *appointment* takes a NOM-NP complement, but no corresponding NOM-INTRANS. The object is obligatory contrary to our defaults, e.g., *John's appointment* must have the interpretation that *John* is the object of *appoint*. Thus, our entry for *appointment* is marked :REQUIRED ((OBJECT)).[8]

4 Our Procedure for Generating Nominalization Patterns

Figure 3 diagrams how nominalization patterns are derived. The rectangles are modules of our algorithm, the ovals are data structures passed between modules and the dotted lines connect the ovals with examples of what the data structures might contain given the sample active clause *IBM appointed Alice Smith.* Due to space limitations, the figure does not include patterns with constituents for the temporal adverbial NP positions (Section 5), e.g., *Mary Smith's June 1, 1998 appointment by IBM* and *Yesterday's appointment of Mary Smith by IBM.* There are an additional five such mappings for *appointee* and an additional twenty-three for *appointment.*

First PET analyzes the sample sentence and identifies the main verb and its arguments (e.g., subject, direct object, etc.). Then it searches NOMLEX for any nominalizations which correspond to the main verb. Our example verb *appoint* has at least two nominalizations: *appointment* and *appointee.* Next the procedure examines the set of :VERB-SUBC classes in each nominalization entry (Figures 1 and 2) and identifies the set of classes which are compatible with the set of arguments in the input. A class is compatible if it allows all the input arguments, and none of its required arguments are missing. For the example sentence, only NOM-NP is chosen for each nominalization. The other phrases all require some phrase in addition to the object, e.g., NOM-NP-AS-NP requires an AS-NP phrase (e.g., *as vice president*). Next the permissible role mappings are generated. By default, the subject and object are optional, but the object is obligatory for *appointment* due to the :REQUIRED feature of its NOMLEX entry. Prepositional phrases are assumed to occur in all orders, so that both (object: of, subject: by) and (subject: by, object: of) are listed in Figure 3. The uniqueness constraint must be obeyed (we cannot have two subjects or two objects). Finally, the syntax of noun phrases only permits the N-N-MOD slot to be filled more than once (*The IBM Alice Smith Appointment*), and in that case our ordering constraint would have to be obeyed. This rules out an interpretation of *The Alice Smith IBM Appointment* where Alice Smith is the appointee and IBM is the appointer.[9]

[7] This ordering constraint is assumed to hold for all nominalizations in English.

[8] Our required/optional settings only apply where some argument is present in a nominalization.

[9] Given a clausal pattern for an example sentence like *They appointed Alice Smith to IBM*, the NOM-NP-PP class would be matched and nominalization patterns would be generated in which IBM (the indirect object)

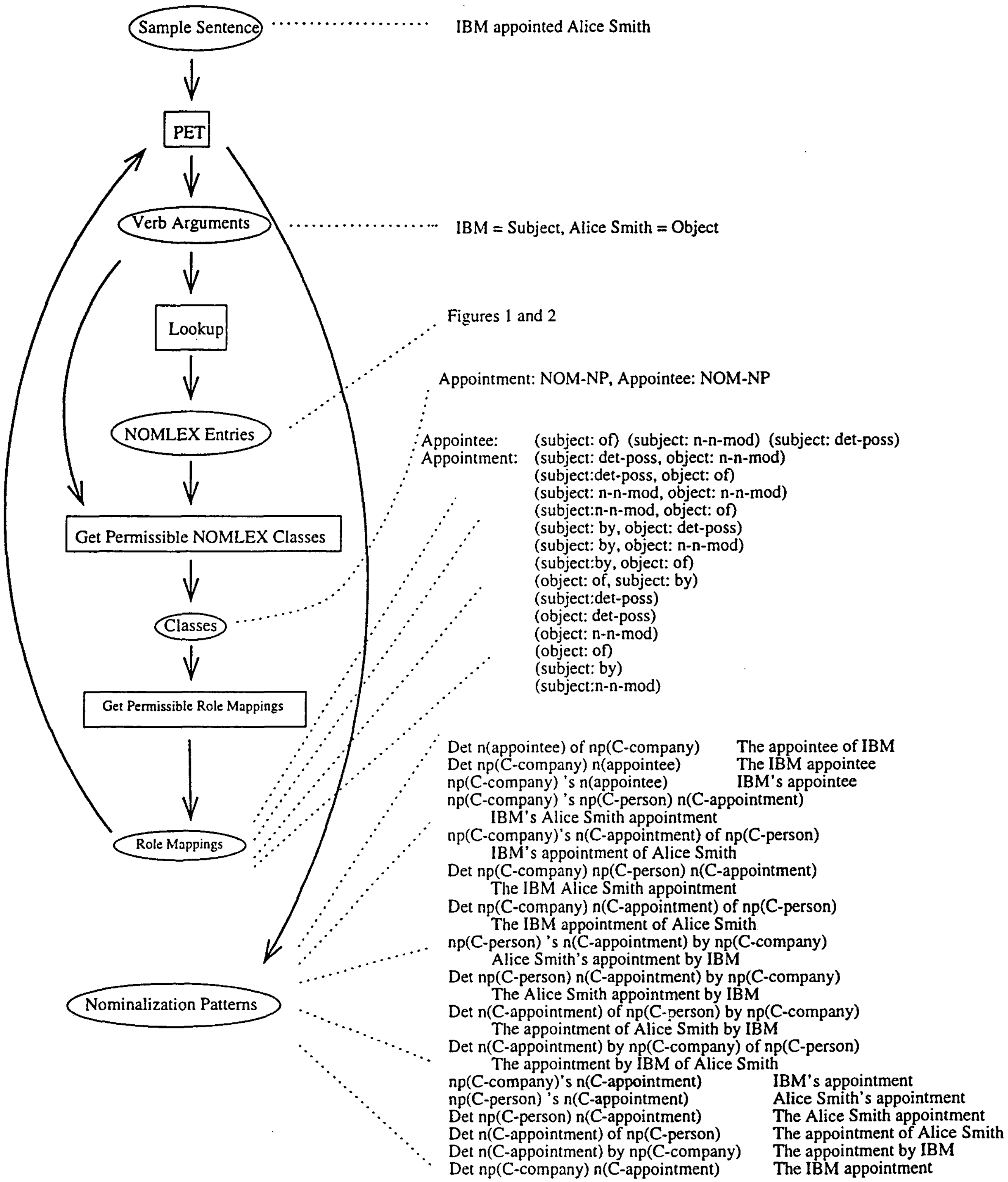
Sample Sentence
IBM appointed Alice Smith
PET
Verb Arguments
IBM = Subject, Alice Smith = Object
Lookup
Figures 1 and 2
NOMLEX Entries
Appointment: NOM-NP, Appointee: NOM-NP
Get Permissible NOMLEX Classes
Classes
Appointee: (subject: of) (subject: n-n-mod) (subject: det-poss)
Appointment: (subject: det-poss, object: n-n-mod)
(subject:det-poss, object: of)
(subject: n-n-mod, object: n-n-mod)
(subject:n-n-mod, object: of)
(subject: by, object: det-poss)
(subject: by, object: n-n-mod)
(subject:by, object: of)
(object: of, subject: by)
(subject:det-poss)
(object: det-poss)
(object: n-n-mod)
(object: of)
(subject: by)
(subject:n-n-mod)
Get Permissible Role Mappings
Role Mappings
Det n(appointee) of np(C-company) The appointee of IBM
Det np(C-company) n(appointee) The IBM appointee
np(C-company) 's n(appointee) IBM's appointee
np(C-company) 's np(C-person) n(C-appointment)
IBM's Alice Smith appointment
np(C-company)'s n(C-appointment) of np(C-person)
IBM's appointment of Alice Smith
Det np(C-company) np(C-person) n(C-appointment)
The IBM Alice Smith appointment
Det np(C-company) n(C-appointment) of np(C-person)
The IBM appointment of Alice Smith
np(C-person) 's n(C-appointment) by np(C-company)
Alice Smith's appointment by IBM
Det np(C-person) n(C-appointment) by np(C-company)
The Alice Smith appointment by IBM
Det n(C-appointment) of np(C-person) by np(C-company)
The appointment of Alice Smith by IBM
Det n(C-appointment) by np(C-company) of np(C-person)
The appointment by IBM of Alice Smith
Nominalization Patterns
np(C-company)'s n(C-appointment) IBM's appointment
np(C-person) 's n(C-appointment) Alice Smith's appointment
Det np(C-person) n(C-appointment) The Alice Smith appointment
Det n(C-appointment) of np(C-person) The appointment of Alice Smith
Det n(C-appointment) by np(C-company) The appointment by IBM
Det np(C-company) n(C-appointment) The IBM appointment

Figure 3: Deriving Nominalization Patterns with PET

PET can then use these mappings to generate patterns, as it does for the various types of clauses. Using pattern matching and dictionary look-up, PET associates the verbal arguments with semantic classes. In our example, the subject is a company and the object is a person. This information can be applied to each mapping to produce a pattern. The nominalization patterns in Figure 3 are generated from the role mappings listed using this semantic information and interpretting the nominal role labels. For example, the mapping (SUBJECT: DET-POSS, OBJECT: PP-OF) generates the nominalization pattern:

np(C-company) 's n(appointment) of np(C-person)

(*IBM's appointment of Alice Smith*).

5 Adjunct Mappings

The preceding section gave a simplified account of mapping nominalization patterns. We must also handle certain adjuncts. Temporal PPs that can occur in clauses can usually occur in nominalizations as well. The positions DET-POSS and N-N-MOD may be occupied by temporal NPs (*Yesterday's appointment of Alice Smith by IBM*, *The January 3, 1998 appointment by IBM of Alice Smith*). When an NP is temporal and occupies either of these positions it may fill a temporal slot in an IE pattern. Since temporal NPs are neither companies, nor people, they will not fill the object or subject slots in the IE patterns above. Therefore, the possibility of filling temporal slots from DET-POSS and N-N-MOD positions should cause no conflicts for *appointment.*

6 Related Work

Other computational linguistics work on decoding nominalizations includes (Hobbs and Grishman, 1976), (Dahl et al., 1987) and (Hull and Gomez, 1996). (Hull and Gomez, 1996) is the most similar to our own in that their ultimate goal is to extract information from the World Book Encyclopedia. That task is defined differently than for our MUC-related work. The lexical entries created by Hull and Gomez include selectional constraints tied to WordNet classes. Their procedure for converting nominalizations into predicate argument structure relies on this semantic information, which they use to distinguish nominalizations from nouns and arguments from adjuncts. Their coverage of arguments is limited to subjects, objects, and prepositional phrases, whereas NOMLEX provides detailed coverage of all core and all oblique arguments.

7 Concluding Remarks

We have been working on NOMLEX for one year with a staff of two part-time graduate students, one full-time staff member and one part-time staff member. We currently have 700 entries and expect to have 1000 entries by Fall of 1998. After testing, we intend to distribute an alpha version of NOMLEX via our FTP site. This paper describes one of the applications for which NOMLEX proves useful. Our hope is that other researchers will apply NOMLEX to new applications. Furthermore, comments from users should prove helpful for updating and revising this resource.

References

Douglas Appelt, Jerry Hobbs, John Bear, David Israel, Megumi Kameyama, Andy Kehler, David Martin, Karen Meyers, and Mabry Tyson. 1995. SRI International FASTUS System: MUC-6 Test Results and Analysis. In *Proceedings of the Sixth Message Understanding Conference.*

Debroah Dahl, Martha Palmer, and Rebecca Passonneau. 1987. Nominalizations in PUNDIT. In *Proceedings of the 25th Annual Meeting of the Association for Computational Linguistics.*

Ralph Grishman. 1995. The NYU system for MUC-6 or where's the syntax? In *Proceedings of the Sixth Message Understanding Conference.*

Jerry R. Hobbs and Ralph Grishman. 1976. The Automatic Transformational Analysis of English Sentences: An Implementation. *International Journal of Computer Mathematics,* 5:267–283.

A. S. Hornby, editor. 1980. *Oxford Advanced Learner's Dictionary of Current English.* Oxford University Press, Oxford.

would follow Alice Smith (the direct object). These patterns would match *The Alice Smith IBM appointment* (or *Alice Smith's IBM appointment*) and give it a very similar interpretation to one in which IBM is the appointer.

Richard D. Hull and Fernando Gomez. 1996. Semantic Interpretation of Nominalizations. In *Proceedings of the Thirteenth National Conference on Artificial Intelligence and the Eighth Innovative Applications of Artificial Intelligence Conference.*

David E. Johnson and Paul M. Postal. 1980. *Arc Pair Grammar.* Princeton University Press, Princeton.

Catherine Macleod, Adam Meyers, Ralph Grishman, Leslie Barrett, and Ruth Reeves. 1997. Designing a Dictionary of Derived Nominals. In *Proceedings of Recent Advances in Natural Language Processing*, Tzigov Chark, Bulgaria.

Catherine Macleod, Ralph Grishman, and Adam Meyers. forthcoming. COMLEX Syntax: A Large Syntactic Dictionary for Natural Language Processing. *Computers and the Humanities.*

1995. *Proceedings of the Sixth Message Understanding Conference.* Morgan Kaufman. (MUC-6).

1998. *Proceedings of the Seventh Message Understanding Conference.* Morgan Kaufman. (MUC-7).

P. Proctor, editor. 1978. *Longman Dictionary of Contemporary English.* Harlow, Essex.

Susanne Rohen Wolff, Catherine Macleod, and Adam Meyers, 1994. *Comlex Word Classes Manual.* Proteus Project, New York University.

Roman Yangarber and Ralph Grishman. 1997. Customization of Information Extraction Systems. In *Proceedings of the International Workshop on Lexically Driven Information Extraction*, Frascati, Italy.