Acquisition of Linguistic Patterns for Knowledge-Based Information Extraction

Sanda M. Harabagiu, Steven J. Maiorano

Department of Computer Science and Engineering
Southern Methodist University
Dallas, TX 75275-0122, U.S.A.
{sanda, steve}@renoir.seas.smu.edu

Abstract

In this paper we present a new method of automatic acquisition of linguistic patterns for Information Extraction, as implemented in the CICERO system. Our approach combines lexico-semantic information available from the WordNet database with collocating data extracted from training corpora. Due to the open-domain nature of the WordNet information and the immediate availability of large collections of texts, our method can be easily ported to open-domain Information Extraction.

1. The Problem

The Message Understanding Conferences (MUCs) and the TIPSTER program gave great impetus to research in Information Extraction (IE). The systems that participated in the MUCs have been quite successful at extracting information from newswire messages and filling templates with the information pertaining to prespecified events of interest. Typically, the templates model queries regarding *who* did *what* to *whom*, *when* and *where*, and sometimes *how*.

During the past years, the performance of some of the IE subtasks has reached near-human precision. For example, current IE systems are capable of recognizing named entities with over 90% precision. The broader task of event recognition (i.e. filling in scenario templates) results in 60% recall and 70% precision, which is comparable with human interannotator agreement ranging between 65% and 80%. These achievements are explained by two factors: (1) usage of the finite-state technology and (2) the availability of domain-dependent knowledge.

The usage of finite-state transducers (Pereira and Wright 1997) has emerged as a dominant technology in the IE field. Finite-state automata, frequently cascaded, are capable of recognizing linguistic constructs ranging from low-level syntactic expressions (e.g. noun groups and verb groups) to higher-level, domain relevant clausal patterns. The recognition of *events and entities of interest* is based on linguistic patterns representing the domain knowledge. Although there have been several remarkable efforts to acquire domain linguistic patterns automatically, some of the most successful systems still employ manually crafted domain rules.

The availability of the WordNet lexico-semantic database (*www.cogsci.princeton.edu/∼wn*) and our previous success in deriving textual implicatures (Harabagiu et al.1996) prompted this research. The extensive semantic net encoded in WordNet can be mined for concepts and lexico-semantic relations relevant to a domain. The resulting concepts and their interrelations are validated by domain- relevant corpora, enabling the discovery of their syntactic contexts. Our novel method generates linguistic patterns for a domain as *production rules* induced when using the principle of maximal coverage of collocating domain concepts.

The rest of the paper is organized as follows. Section 2 contrasts several other methods for automatic acquisition of linguistic patterns with our approach. Section 3 details the representation of domain semantic spaces and the techniques used to create and mine them from online resources. Section 4 reports and discusses the experimental results and Section 5 summarizes the conclusions.

2. Acquisition of linguistic patterns

The portability of IE systems across different domains is hindered by the need of new linguistic patterns for each novel domain. The development of regular patterns for a new domain is time-consuming and relies on computational linguistic expertise (e.g. porting BBN's PLUM systems from the Joint Venture domain (MUC-5)(MUC-5) to the Microelectronics domain (MUC-5) (MUC-5) took 3 person weeks). Many systems, including those from NYU (Grishman 1995), BBN (Weishedel 1995), SRI (Appelt et al.1995), SRA (Krupka 1995), MITRE (Aberdeen et al.1995) and University of Massachusetts (Fisher et al.1995) have taken steps to simplify the acquisition of domain-specific patterns.

Researchers have considered two different ways of tackling the domain knowledge problem:
(1) the automatic acquisition of linguistic patterns from training corpora and
(2) the development of a specification language that allows the developer to write regular productions in the most economical way.
The latter endeavor is illustrated by FASTSPEC (Appelt et al.1995), a linguistic pattern specification language developed at SRI International that translates any regular expression into a finite state machine using an optimizer compiler. The acquisition of domain patterns was considered in many systems, starting with Auto-Slog (Riloff 1993), PALKA (Kim and Moldovan 1995) and CRYSTAL (Soderland et al.1995). All these systems learn linguistic patterns in the format of semantic case frames using different learning strategies. In addition they all rely on pre-processed domain knowledge:

• annotated texts corpora (e.g. in the case of Auto-Slog[1]

[1] Auto-Slog was later enhanced into Auto-Slog-TS for which texts do not require annotations, but mere classifications ac-

all noun phrases that should be extracted are annotated in the texts) or

• manually constructed concept hierarchies (e.g. for CRYSTAL or PALKA). These hierarchies also encode selectional constraints.

WordNet was used recently in two linguistic pattern acquisition systems. Califf and Mooney's RAPIER system (Califf and Mooney 1997) learns unbounded ELIZA-like patterns (Weizenbaum 1966) by using limited syntactic information (i.e. the output of Brill's part-of-speech tagger (Brill 1992)) and semantic class information from WordNet. RAPIER was used to extract information regarding job offerings, a domain that does not include scenario templates of MUC-like complexity. Unlike any of the other systems, RAPIER learns rules that specify constraints at the word level rather than at the constituent level.

The second study, presented in (Bagga et al.1997), assesses the role of WordNet in learning general patterns by using subsumption operators along the noun hierarchies of WordNet. In this pattern acquisition system, the semantic ambiguity of the head-words is resolved by human intervention. Moreover, similar to PALKA, initial patterns are entered by a user (with the help of a GUI inspired by SRA's HASTEN system (Krupka 1995)).

Our experiments in domain-knowledge acquisition for IE using WordNet, built upon all these previous systems. In our approach we:

1. minimize human intervention by enhancing WordNet with knowledge imposed by the pattern acquisition process;
2. produce patterns in the FASTSPEC language, thus granting optimal regular productions; and
3. develop an acquisition methodology that does not require semantic disambiguation of the trigger words.

Moreover, we aimed to produce a semantic space of WordNet concepts and relations that can be used to recognize relevant linguistic expressions from any text corpora, even from the Internet.

3. Domain Knowledge for IE

Our method of acquiring linguistic patterns was devised as a three step process:

Step 1. First, we create a semantic space that models the domain via WordNet concepts and relevant connections between them. Building a semantic space for a domain of interest provides means for (a) finding all linguistic patterns that cover the relevant textual information in documents and moreover (b) enables the interpretation of the interrelations between different relevant textual expressions from the same document or even across documents (i.e. document threading).

A semantic space corresponding to a certain domain contributes to the resolution of some of the problems that still hinder the performance of current IE systems:

1. event recognition (also known as template merging),
2. the inference of implicit arguments, and
3. the interpretation of non-literal textual expressions of relevance to a given domain.

Step 2. In the second phase, we scan the phrasal parses of texts from the MUC corpora for collocating domain concepts that are connected in the domain semantic space. Production rules are induced using the principle of maximal coverage of collocating concepts. The phrasal parser implemented in our CICERO IE system generates correct syntactic links emerging from domain concepts, and thus enables the derivation of FASTSPEC-like linguistic patterns.

Step3. Finally, the patterns are classified against the WordNet hierarchies and only the most general linguistic domain patterns are retained. The results matched all the linguistic patterns hand-crafted previously for CICERO and produced several new patterns.

Similar to other knowledge-based tasks, this method of automatically acquiring domain linguistic patterns has had a large start-up effort. This included the need for an unsupervised method of encoding morphological links in WordNet as well as heuristics for reformatting the information from conceptual definitions. However, the high performance of this linguistic pattern-acquisition method indicates that it is a valuable tool for building ontologies of domain patterns, and extremely useful for indexing digital libraries as well.

3.1. Initial experiments

Our initial experiments of mining the WordNet lexical database for domain-dependent lexico-semantic information revealed two interesting facts. First, we expected to derive patterns of interrelated lexical concepts from the large WordNet semantic net and to obtain a high recall. However, we noticed that concepts relevant for a variety of domains were encoded in WordNet, and there were *indirect* relations among them. Thus we learned that linguistic patterns can be retrieved, but with the price of semantic processing.

In our quest for retrieving linguistic patterns directly from the WordNet semantic network, we experimented with three domains: (1) the *management succession* events, tested during the MUC-6 competition; (2) *aircraft crashes*, used for training in MUC-7; and (3) corporation *joint ventures*, tested in MUC-5. For each of these domains we relied on the scenario template keys to identify the domain concepts in the texts. In addition, we employed the MINIPAR named entity recognizer (Lin 1994) to distinguish the names of persons, organizations and locations.

The definition of the templates determines some basic relations that must be uncovered by the linguistic patterns. For example, in the case of management succession events, the template contains slots for the person name; for the management position; and for the organization name. In our experiments, for each template key corresponding to an event, we identified the text fragment containing all the slot values of the template. Next we parsed the text fragments with a high precision parser (Collins 1996) to identify verbs or nominalizations syntactically connected to the key values. At the end, each connection from a verb/nominalization to a key value was searched in WordNet. The search took into account all possible semantic senses of the retrieved verbs and of the non-named entities.

cording to the relevance to the domain of interest. Auto-Slog-TS instead relies on a set of heuristics derived from the patterns induced by Auto-Slog for that domain. In a final stage, patterns are statistically filtered.

Named entities were replaced by the class they represent. The results showed that less than 8% of the relations were accounted for in WordNet.

For example, the following template key, represents a management succession event extracted from the *Wall Street Journal* document with ID 9404250043. We were able to identify the text fragment where the event is described by retrieving the sequence of sentences that contain any of the slot values of the template. The template is:

```
<SUCCESSION_EVENT-9404250043-1> :=
    SUCCESSION_ORG: <ORGANIZATION-9404250043-1>
    POST: "chairman"
    IN_AND_OUT: <IN_AND_OUT-9404250043-1>
                <IN_AND_OUT-9404250043-2>
    VACANCY_REASON: DEPART_WORKFORCE
    COMMENT: "Purdum out, Haley in as chmn of Armco"
<ORGANIZATION-9404250043-1> :=
    ORG_NAME: "Armco Inc."
    ORG_ALIAS: "Armco"
    ORG_DESCRIPTOR: "Steelmaker"
   / "The specialty steelmaker"
    ORG_TYPE: COMPANY
<IN_AND_OUT-9404250043-1> :=
    IO_PERSON: <PERSON-9404250043-2>
    NEW_STATUS: OUT
    ON_THE_JOB: NO
    COMMENT: "Purdum out"
           / "ON_THE_JOB: 'retired'"
<IN_AND_OUT-9404250043-2> :=
    IO_PERSON: <PERSON-9404250043-1>
    NEW_STATUS: IN
    ON_THE_JOB: UNCLEAR
    OTHER_ORG:  / <ORGANIZATION-9404250043-2>
    REL_OTHER_ORG:  / OUTSIDE_ORG
    COMMENT: "Haley in -- came from differ-
ent org, but this may be nonrelevant since he re-
tired from there and has since then been on the Armco
 board"
<PERSON-9404250043-1> :=
    PER_NAME: "John C. Haley"
    PER_ALIAS: "Haley"
    PER_TITLE: "Mr."
<PERSON-9404250043-2> :=
    PER_NAME: "Robert L. Purdum"
    PER_ALIAS: "Purdum"
    PER_TITLE: "Mr."
```

The corresponding retrieved text fragment is:
At its annual meeting, **Armco** *also named* **John C. Haley,** *64 years old, chairman.* **Mr. Haley's** *appointment is for a one-year term, during which* **Armco's** *board will study the concept of a nonexecutive* **chairman. Mr. Haley**, *an* **Armco** *board member since 1975, is retired* **chairman** *and chief executive officer of closely held Business International Corp. He succeeds* **Robert L. Purdum,** *58, who retired.*

The parser, in conjunction with the named entity recognizer, identified the following SVO patterns from the above text fragment:

"Armco[ORGANIZATION] named John Haley[PERSON] chairman[POSITION]"
Pattern 1: [*Subject*=Organization][*Verb*=name] [*Object1*=Person][*Object2*=Position]

Mr. Haley[PERSON] is retired chairman[POSITION] of Business International Corp.[ORGANIZATION]"
Pattern 2: [*Subject*=Person][*Verb*=be][*Object1*=Position] [*Preposition*={of}][*Preposition-object*=Organization]

In WordNet 1.6 there is no relation between *organization* and any of the senses of the verb *name*, nor between *name* and any concept subsumed by *position*. Similarly, for pattern 2, no connections between its concepts could be found in WordNet.

However we also noticed that WordNet encodes a wealth of verbal concepts. Since each domain is characterized by a number of lexicalizations of the most relevant events, actions or states, we can access relations among these verbal concepts. For example, in the MUC-6 domain, verbs such as *fire* and *hire* are directly connected as antonyms. Many more domain verbs are connected, albeit indirectly. The latter category of verbs has the property that they share many common concepts in their defining glosses. For example, the verb *appoint*, with the semantic sense *2* in WordNet is defined as *assigning a position*, but is not related to any of the senses of the verb *assume*. However, the relationship between *appoint* and *assume* is transcended by the relation between *giving* and *taking*, already encoded in WordNet as an entailment.

In a second set of experiments, we have found that if we start with a predefined set of linguistic rules, expressed as subject-verb-object (SVO) patterns, the above observations help enhance the set with additional rules. Thus we noticed that novel connections between domain concepts do not result directly from available WordNet relations, but they are rather combinations of WordNet relations mixed with :

- (i) *lexico-semantic relations* implicit in the conceptual definitions (known as *glosses* in WordNet);
- (ii) *morphologically cued relations*;
- (iii) *concept-identity relations* established between a synset and all its usages in the gloss of other synsets; and
- (iv) *collocational relations*, connecting multi-word synset elements with the synsets of each word[2](e.g. synset {*take office*} has collocating relations to synsets {*fill, take*} and {*office, position, post, berth, spot, place, situation*}).

Our general conclusion after these experiments was that although WordNet displays a magnitude of linguistic information, acquiring domain knowledge for IE involves a complex search and the derivation of several additional forms of connections among concepts. For example, a new pattern for the MUC-6 domain was found because pf the connection between the trigger words *take* and *helm* (as a form of position of leadership). The FASTSPEC representation of this new pattern is:

[*Subject*=Person][*Trigger-phrase*=take the helm] [*Preposition*={at|of}][*Preposition-object*=Organization]

This pattern extends the general SVO structure of the linguistic patterns implemented in IE systems, allowing more complex triggers. The acquisition of this pattern is derived by the WordNet relations between synsets {*assume, take on, take over*} and {*position, office, place, post, slot*}. Synset {*take office*} can be reached via:
(a) *concept-identity* relations, since the concepts *assume* and *office* are both used in the gloss of *take office*.
(b) a *collocational* relation, generated by the same sense of *office* in the synsets {*position, office, place, post, slot*} and {*take office*}.
Moreover, synset {*take office*} is used to define synset {*accede to, enter upon*}, a hyponym of {*succeed, come after, follow*}. Therefore we infer that a succession event can

[2]These relations implement the assumptions of compositional semantics.

be expressed also by any collocation of the verb *take* (with the semantic sense from *take office*) and

1. any element from the synset {*position, office, place, post, slot*};
2. any of its hypernyms; or
3. any synset defined[3] in the hierarchy of {*position, office, place, post, slot*}.

A synset that pertains to case (c) is {*helm*}, defined as *(position of leadership)*. Therefore, *[take the helm]* is induced as a novel trigger phrase.

Learning new patterns involves not only deriving new trigger words, but also their satellites (e.g. *Subject*, *Object*, or *Prepositional Object*). Collecting all the collocations of trigger words from a corpus is not sufficient for establishing meaningful connections in a domain. Thus, we need to validate the connections of the trigger concepts in a semantic space that models the domain of interest. For example, in finding the satellites of trigger-phrase *take the helm*, we searched for connections to *management-position* or *manager*. WordNet provides a connection between synsets {*helm*} and {*manager, director, managing director*}. The gloss of {*helm*} defines it as a *position* having the attribute of *leadership*. In turn concept *leadership* is a nominalization of the verb *to lead*. Another nominalization of the verb *lead* is *leader*, the subject of the action. Because synset {*leader*} is a hypernym of *manager*, a semantic connection between *helm* and *manager* is found. This indicates possible pattern matchings of *taking the helm* and any position of management.

We conclude that at the core of these experiments is the construction of domain semantic spaces, encoding several additional forms of relations, detailed in Section 3.2..

3.2. Building semantic spaces for IE

Domain knowledge is acquired in the form of a semantic space formalized as a triplet

< **concepts, connections, contexts** >

The set of **concepts** is represented by a collection of WordNet synsets found elevant to a given domain. The **connections**, spanning the semantic space concepts , model several forms of relationships:

1. *Thematic connections* between concepts in a certain *context*. Thematic relations are derived from (a) lexico-semantic relations encountered in the gloss definitions; (b) morphological relations; and (c) interpretation of WordNet paths, glosses and morphological relations. Figure 1 illustrates the derivation of the matic relations for the three cases.
2. *Subsumption connections*, generated either from original WordNet *IS-A* relations or from the interpretation of gloss geni. Figure 2 illustrates subsumption connections typical of the `MUC-6` domain. Concept `PERSON` is subsumed by a sequence of WordNet synsets, connected by *ISA* relations. Concept `POSITION` subsumes synset {*headship*}, whose gloss is linked via a *concept-identity* relation to synset {*head, chief, top dog*}. Consequently, `POSITION` also subsumes the hyponyms of *head*.

[3](i.e. having the genus of the gloss)

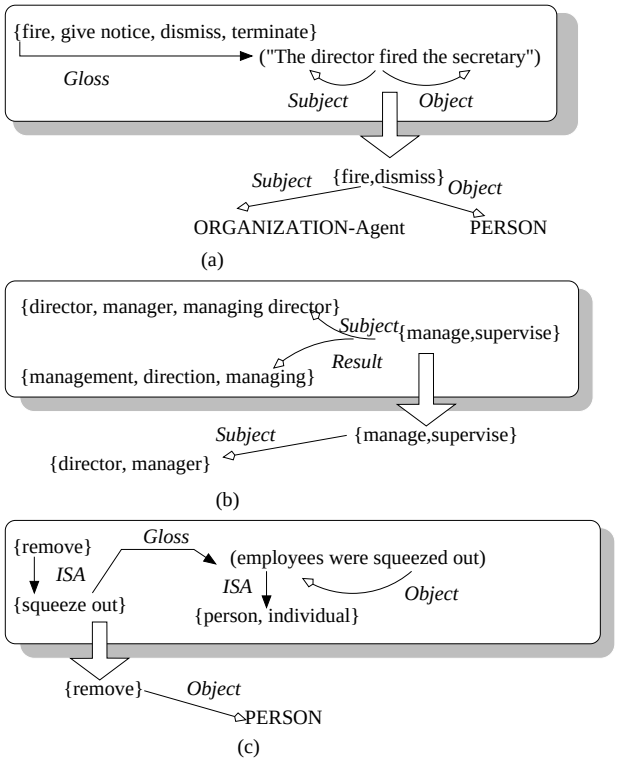

Figure 1: Thematic connections derived from WordNet

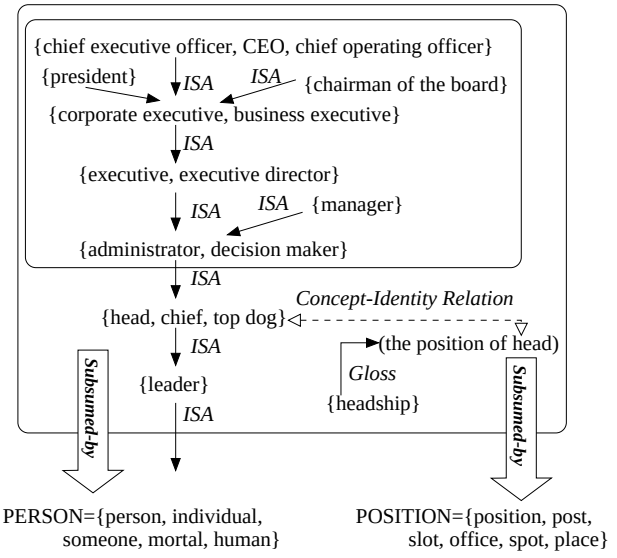

Figure 2: Subsumption connections in the MUC-6 domain

3. *Contextual connections* spanning the context objects and describing the possible relationships between them. We distinguish four types of contextual connections: *entail* and *antonym* connections, encoded in WordNet and *compose* and *similar* connections. We assume that a contextual object *entails* another one if all propositions true in the former will remain true in the latter. A context is *antonymous* to another if any of the propositions that hold in its space will not be true in the latter, and vice versa. Assuming that a proposition P_1 holds in context C_1 and a proposition P_2 holds in a context C_2, if there is a context C_0 in which both P_1 and P_2 hold, we say that there are *compose* connections from C_0 to both C_1 and C_2. Finally, when

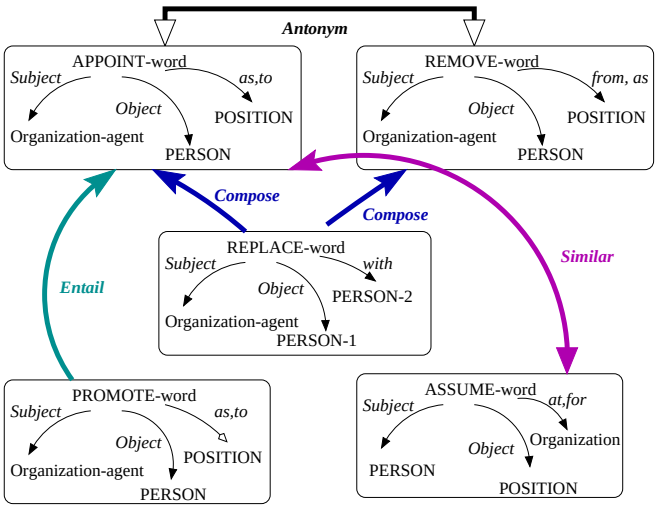

Figure 3: Contextual connections in the MUC-6 domain

all propositions holding in a context C_1 hold also in C_2 (and vice versa), we establish a *similar* connection between C_1 and C_2. Figure 3 illustrates several contextual connections typical for the MUC-6 domain.

Contexts are semantic objects encompassing : (a) concepts, (b) thematic and subsumption connections and (c) conditions that enable their inter-connections. Contexts model various ways of expressing information regarding events of interest in a given domain, and are a means of capturing the relationship between these events. For example, in the MUC-6 domain, the event of appointing a person to a new managerial position can be expressed by stating that the respective person has been promoted, or by announcing the person's new position, or by stating that the person became or is the executive in that position, or by stating that the person stepped into the new position. Since promoting (or becoming, stepping in or being) cannot always be viewed as a form of appointing, we consider *entailment* (or implication) connections between these events (modeled by different contexts). Figure 4 illustrates an example of the general conditions under which connections between contexts hold.

APPOINT-context

Similar to: Step-in Context Conditions: person(appoint)=person(step-in) position(appoint)=position(step-in) Org-agent(appoint)=Org-agent(step-in)
Composed by: Replace Context Conditions: person(appoint)=person2(replace) position(appoint)=position(replace) Org-agent(appoint)=Org-agent(replace)

Figure 4: Conditions set on contextual connections: An example

This model of contextual objects based on knowledge from WordNet is in accordance with the formal theories of contexts reported in (McCarthy 1993) and (Buvač 1996).

When the domain of interest is defined as a natural language text, with a structured sequence of words, (e.g. a complex nominal =*management succession* in the case of MUC-6 or *aircraft crashes* for the dry-run of MUC-7)[4], the semantic space is acquired by the following algorithm:

Algorithm 1 (Builds semantic spaces)
Input: Sequence of words: $\{word_k\}$
Output: <*concepts, connections, contexts*>[5]

◇ *Procedure:*
For every *word* defining the domain
1. Retrieve all the morphological variations and the WordNet synsets containing the *word*.
2. Classify the synsets according to maximal coverage of relations between concepts.
3. Search for common concepts related to synset classes corresponding to different domain words.
4. If no common concepts are found then
   goto 5. else
   4.1. Discard all other classes
   4.2. Derive thematic and subsumption connections
5. Expand each synset class along:
   (a) *IS-A* links (and *antonyms* in the case of collocational synset entries)
   (b) WordNet meronymic relations (e.g. *is-member* or *is-part*)
   (c) synsets that contain collocations of concepts from that class
   (d) gloss geni not contained in the class
   (e) gloss concepts densely connected in the class
6. Repeat steps *2.–4.* until only one semantic class corresponds to each *word*.
7. Derive contexts (and subsumption and thematic connections).
8. Specialize every context in the domain by
   8.1. taking all hyponyms having subjects, objects or prepositional relations to common concepts
   8.2. retrieving concepts that have the common concepts in their glosses
9. Generalize all classes of events or entities in every context.
10. Derive contextual connections.

3.3. Walk-Through Example

The first step of acquiring a semantic space for the MUC-6 domain described by the complex nominal *"management succession"* produces morphological derivations as a side effect of classification in clusters of highly related synonyms sets. We have studied the heuristics that connect synsets containing words produced by derivational morphology in WordNet 1.6. We have found that words having the same morphological root, but different parts-of-speech tend to be interrelated by one of several functions. For example, nouns derived from a verb (i.e. nominalizations) may
(a) reflect the function of the action itself, thus are *act-nominalizatons*, or

[4]The methodology can be extended easily when the domain is defined by a list of keywords or by several free text paragraphs, as was the case for TREC-6 or TREC-7

[5]A semantic space.

(b) represent the agents of the action lexicalized by the verb, thus are *subject-nominalizations* or
(c) represent the effects of the action, thus are *result-nominalizations*.

Examples of heuristics that establish the function of the morphological derivation, and hence an interrelation between the corresponding synsets are:
Heuristic 1 A noun synset S_N is an *act-nominalization* of a verb synset S_V when the gloss of S_N has a genus of the form "*the action of*<*verb_s*>, where *verb_s* is a member of S_V.
An example is rendered by sense 3 of the nominalization *succession*, which is the act-nominalization of sense 2 of the verb *succeed*.
Heuristic 2 A noun synset S_N is a *subject-nominalization* of a verb synset S_V if the genus of the gloss of S_N (or of one of its hypernyms) is the subject of *verb_s*, where *verb_s* is an element of S_V or of any of its hypernyms.
An example is sense 1 of the noun *manager*, whose hypernym {*administrator, decision maker*} has the gloss *(someone who administers a business)*, with the verb *administer*, belonging to the same WordNet hierarchy as sense 2 and 4 of the verb *manage*.
Heuristic 3 A noun synset S_N is a *subject-nominalization* of a verb synset S_V if in the hierarchy of S_N we find another *subject-nominalization* of a verb from the same hierarchy as S_V.
This heuristic makes sense 1 of *manager* a *subject-nominalization* of senses 2 and 4 of the verb *manage* because we have *administrator* from the hierarchy of *manager*#1 a *subject-nominalization* of *administer* from the hierarchy of *manage*#2. Each of these heuristics retrieves morphological relations, producing the results of the first step of Algorithm 1.

A group of synsets linked by morphological relations define a class of synsets. They represent the output of the second step of Algorithm 1. A remarkable side-effect of the application of these heuristics is the clustering of WordNet semantic senses. One of the main criticism of WordNet is its fine granularity of the definition of semantic senses. From the definitions of the WordNet senses for all morphological variations of *management* and *succession* we find that in the MUC-6 domain both senses 1 and 4 or the noun *succession* are applicable to the changes made in management position. Similar senses 2 and 4 of the verb *manage* describe the activity of a person holding a managerial position. They represent the results of the second step of Algorithm 1.

In the third step of Algorithm 1, we seek common concepts related to any of classes for *management* or *succession*. The search is performed along:
(1) collocation entries containing words from the hierarchies of both words;
(2) the genus hierarchies; or
(3) along the concepts directly related to the geni of the conceptual glosses (usually via *subject*, *object* or *prepositional* relations.)
This search produced several common concepts between *Class 1-management* and *Class 1-succession*: the concepts *take office* and *leave office*. Finding concepts connected to both *management* and *succession* produces also a shallow interpretation of the complex nominal *"management succession"*. Because the common concepts were found as special cases (i.e. hyponyms) of the actions lexicalized by the verb *succeed*, the meaning of the complex nominal becomes the event of succeeding by taking or leaving positions of management. This interpretation generates the first contextual object of the semantic space: the *Succession-Context*.

Following the discovery of the common concepts at step 4.1, all the other classes of synsets are discarded. Step 4.2. of Algorithm 1 derives the subsumption and thematic connections. The thematic connections for the *Succeed-Context* are derived from sense 2 of *successor* and from the gloss of {*position, office*}, a synset collocating in both *take office* and *leave office*. Thematic connections are propagated along the subsumption chains.

Step 5 of Algorithm 1 expands the semantic space with novel domain concepts, which populate the contextual objects derived at step 7. Some of conceptual objects derived from the initial common concepts are : the {*take office, assume*} context and the {*step down, leave office*} context. The antonymy WordNet relation between these two concepts translates into distinct *Person* fillers. Concepts subsumed by {*step down, leave office*} in the *Leave office-Context* are connected to concepts from the *Remove-Context*. In WordNet 1.6, the sense of the verb *retire* subsumed by *leave office* is also *entailed* by another sense of the verb *retire*, whose hypernym is {*remove*}. These relations translate into an *Entailment* connection between the two contexts.

The verb *fire*, subsumed by *remove* has an antonym: verb *hire*. The latter concept is subsumed by verb *appoint*. Both the *Appoint-Context* and the *Remove-Context* are linked through *Compose-connections* to the *Replace-Context*. The latter context was inferred from the codification in WordNet of two senses of the verb *replace*: one that subsumes verb *succeed*, the other one entailing it. Consequently, the *Succeed-Context* and the *Replace-Context* are similar as they refer to the same event, but employ different themes.

The inference of these conceptual connections takes place in step 10 of Algorithm 1. In steps 8 and 9 specializations and generalizations take place in every context, as a side effect of the inference of subsumers.

Further specializations and generalizations of each context are performed in an empirical way, by combining the hierarchical organization of WordNet with the information provided by the MUC-6 text corpus.

4. Empirical Analysis

The methodology of creating a semantic space needs to be validated by a corpus-based empirical test. Table 1 lists the number of domain concepts, contextual objects, subsumption and thematic connections as well as the number of contextual connections obtained for the MUC-6 domain.

Three problems arise when using the domain contextual objects to devise linguistic patterns:

1. As WordNet synsets are not encoded for a specific domain, many of the synsets gathered in the contextual

Nr. words	Nr. concepts	Nr. contextual objects	Nr. subsumption connections	Nr. thematic connections	Nr. contextual connections
245	81	20	45	104	32

Table 1: Cardinality of the semantic space built for MUC-6

objects contain entries that are not used in the respective domain.

2. Thematic relations were induced only from the conceptual glosses. Text describing events from a given domain generally display far more thematic relations than those from the definitions of concepts. These relations should be incorporated in the linguistic rules.

3. The degree of generality of concepts from every context has to be done in harmony with the generality of the concepts employed in real world texts for the domain of interest.

These problems are resolved as a by-product of a corpus-based procedure that acquires the domain linguistic patterns.

Algorithm 2 (Finds domain linguistic patterns)
Input: Contexts from the semantic space, Text corpora.
Output: Linguistic rules.

Procedure:
1. For every *Contextual object* from the semantic space of the domain
2. For every $\mathcal{V}$ *verb* or *Act-nominalization*
3. Scan all texts and gather the phrasal context where $\mathcal{V}$ or any concept subsumed by it occur
4. If (there is a phrasal context where a new thematic role for $\mathcal{V}$ exists)
5. If (all the other roles of $\mathcal{V}$ are encountered in that phrasal context as well)
6. Create a new contextual object for $\mathcal{V}$.
7. If (the filler of the new role subsumes any of the existing fillers)
8. Add the new prepositional-role for that filler.
9. For every *Contextual object* from the semantic space of the domain
10. Find the most general filler.
11. Find the synset elements that were retrieved from phrasal contexts.
12. Create a linguistic rule and mark its label with *RULE-label*.
13. Mark the verbal concepts encountered in text with the *RULE-label*-$\mathcal{V}$ attribute.
14. Mark the thematic filler words encountered in text with the *RULE-label*<*theme*> attribute.
15. Translate the themes in FASTSPEC.

Table 2 illustrates the results of this procedure applied to the MUC-6 domain. The MUC-6 corpus was preprocessed with the CICERO phrasal parser. We have devised only one novel context (and consequently a new rule) since promote, a subsumer of appoint was found to have a supplementary theme provided by the previous position of the promoted executive. Moreover, we have automatically produced all the linguistic patterns that were manually crafted for CICERO and came up with several novel linguistic rules, corresponding to the *Step-down* and *Take-the-helm* contexts. In addition, by combining the knowledge from WordNet with the experience of building CICERO we have devised a methodology that creates rapidly and easily linguistic patterns for any new domain. The existence of the semantic space (and its contextual connections) provides with the relational semantics between the events extracted from texts, and makes possible event correference (or merging) and threading across documents. We contemplate the usage of the semantic space for information retrieval from the Internet and to the task of summarization.

Nr. rules	Nr. thematic connections	Nr. words encountered in texts
21	209	193

Table 2: Attributes of linguistic rules derived for MUC-6

5. Conclusions

This paper describes an implementation of a system that acquires domain linguistics rules for IE with the use of the WordNet lexico-semantic database. Two different algorithms that participate in the acquisition are presented. The first algorithm generates a semantic space for each domain. This semantic space is an important resource that can be used for other aspects of the IE process as well. For example, the event merging, i.e. the process of recognizing the same event in a text, could greately benefit from such a resource. Similarly, coreference resolution in IE systems can be enhanced when a semantic space is available.

For the particular case of linguistic pattern acquisition, we showed that this method generates very rich semantic spaces that are used by the second algorithm to generate set of linguistic patterns. This methodology contrasts with recent approaches that use machine learning to simultaneously learn patterns and dictionaries for IE, as presented in (Riloff and Jones 1999). Mining a large dictionary like WordNet proves to bring forward very useful domain knowledge. In the future, we plan to integrate this approach with bootstrapping techniques, similar to those reported in (Riloff and Jones 1999).

6. References

John Aberdeen, John Burger, David Day, Lynette Hirshman, David D. Palmer, Patricia Robinson and Marc Vilain. Description of the ALEMBIC System Used for MUC-6. In *Proceedings of the Sixth Message Understanding Conference (MUC-6)*, pages 141–155. San Mateo, Morgan Kaufmann, 1995.

Douglas E. Appelt, Jerry R. Hobbs, John Bear, David Israel, Megumi Kameyama, and Mabry Tyson. The SRI MUC-5 JV-FASTUS Information Extraction System. In *Proceedings of the Fifth Message Understanding Conference (MUC-5)*, 1993.

Douglas E. Appelt, Jerry R. Hobbs, John Bear, David Israel, Megumi Kameyama, Andrew Kehler, David Martin, Karen

Myers and Mabry Tyson. Description of the FASTUS System Used for MUC-6. In *Proceedings of the Sixth Message Understanding Conference (MUC-6)*, pages 237–248. San Mateo, Morgan Kaufmann, 1995.

Amit Bagga, Joyce Yue Chai and Alan Biermann. The Role of WordNet in The Creation of a Trainable Message Undestanding System. In *Proceedings of the 14th Conference on Artificial Intelligence (AAAI/IAAI-97)*, 941–948.

John Bear and Jerry R. Hobbs. Localizing expression of ambiguity. In *Proceedings of the Second Conference on Applied Natural Language Processing, Association for Computational Linguistics*, pages 235–242, 1988.

Richard Bobrow, Ron Ingria and David Stallard. The Mapping Unit Approach to Subcategorization. In *Proceedings of the Speech and Natural Language Workshop*, 1991.

Eric Brill. A simple rule-based part of speech tagger. In *Proceedings of the Third Conference on Applied Natural Language Processing*, pages 152–155, 1992.

Sasa Buvač. Quantificational logic of context. In *Proceedings of the 13th National Conference on Artificial Intelligence (AAAI-96)*, pages 600–606, Portland, OR, 1996.

Mary Elaine Califf and Raymond J. Mooney. Relational Learning of Pattern-Match Rules for Information Extraction. In *Proceedings of the ACL Workshop on Natural Language Learning*, pages 9–15, 1997.

Michael Collins. A New Statistical Parser Based on Bigram Lexical Dependencies. In *Proceedings of the 34st Annual Meeting of the Association for Computational Linguistics, ACL-96*, pages 184–191, 1996.

David Fisher, S. Soderland, J. McCarthy, F. Feng and Wendy Lehnert. Description of the UMass System Used for MUC-6. In *Proceedings of the Sixth Message Understanding Conference (MUC-6)*, pages 127–140. San Mateo, Morgan Kaufmann, 1995.

Paul J. Grice. Logic and conversation. In P. Cole and J. Morgan, editors, *Syntax and Semantics Vol.3: Speech Acts*, pages 41–58. Academic Press, New York, 1975.

Ralph Grishaman. New York University PROTEUS System: MUC-4 Test Results and Analysis. In *Proceedings of the Sixth Message Understanding Conference (MUC-6)*, pages 124–127. San Mateo, Morgan Kaufmann, 1995.

Sanda Harabagiu, Dan Moldovan and Takashi Yukawa. Testing Gricean constraints on a WordNet-based Coherence Evaluation System. In *Working Notes of the AAAI-96 Spring Symposium on Computational Approaches to Interpreting and Generating Conversational Implicature*, pages 31–38, 1996.

Jerry R. Hobbs, Douglas E. Appelt, John Bear, David Israel, Megumi Kameyama, Mark Stickel and Mabry Tyson. FASTUS: A Cascaded Finite-State Transducer for Extracting Information from Natural-Language Text. In *Finite State Language Processing*, edited by Emmanuel Roche and Yves Schabes, MIT Press, 1997.

Jun-Tae Kim and Dan I. Moldovan. Acquisition of Linguistic Patterns for Knowledge-Based Information Extraction. In *IEEE Transactions on Knowledge and Data Engineering*, 75(5):713–724, 1995.

George Krupka. Description of the SRA System as Used fro the MUC-6. In *Proceedings of the Sixth Message Understanding Conference (MUC-6)*, pages 221–235. San Mateo, Morgan Kaufmann, 1995.

Dekang Lin. Principar - an efficient, broad coverage, principle-based parser. In *Proceedings of the 15th International Conference on Computational Linguistics, COLING-94*, pages 482–488, Kyoto, Japan, 1994.

John McCarthy. Notes on formalizing context. In *Proceedings of the 13th International Joint Conference on Artificial Intelligence (IJCAI-93)*, pages 555–560, 1993.

George A. Miller. WordNet: a lexical database for English. In *Communications of the ACM*, Vol.38, No.11:39–41, 1995.

Proceedings of the Fourth Message Understnading Conference (MUC-4). San Mateo. Morgan Kaufmann, 1992.

Proceedings of the Fifth Message Understnading Conference (MUC-5). San Mateo. Morgan Kaufmann, 1993.

Proceedings of the Sixth Message Understnading Conference (MUC-6). San Francisco. Morgan Kaufmann, 1995.

Fernando Pereira and Rebecca Wright. Finite-state approximation of phrase-structure grammars. In *Finite State Language Processing*, edited by Emmanuel Roche and Yves Schabes, MIT Press, 149–179, 1997.

Ellen Riloff. Automatically Constructing a Dictionary for Information-Extraction Tasks. In *Proceedings of the 11th National Conference on Artificial Intelligence (AAAI-93)*, 1044–1049.

Ellen Riloff and Rosie Jones. Learning dcitionaries for Information Extraction by multi-level bootstrapping. In *Proceedings of the 16th National Conference on Artificial Intelligence (AAAI-99)*.

Stephen Soderland, David Fisher, Jonathan Aseltine and Wendy Lehnert. CRYSTAL: Inducing a Conceptual Dictioanry. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI-95)*, Montreal, Canada, 1314–1319, 1995.

Beth Sundheim. Overview of the MUC-6 evaluaiton. Proceedings of the Sixth Message Understnading Conference (MUC-6). San Francisco. Morgan Kaufmann, pages 13–32, 1995.

Ralph Weischedel. Description of the PLUM System as Used for MUC-6. In *Proceedings of the Sixth Message Understanding Conference (MUC-6)*, pages 55–69. San Mateo, Morgan Kaufmann, 1995.

J. Weizenbaum. ELIZA - A computer program for the study of natural language communicaiton between men and machines. In *Communications of the ACM*, No.9:36–45, 1966.