Using Statistical Parsers and Wordnet Ontology for Building Semantic Structures from Encyclopedic Texts

UCF-CS-TR-13-06
Fernando Gomez
Department of EECS
University of Central Florida
Orlando, Fl 32861

Abstract

Algorithms are described for constructing semantic structures from encyclopedic texts and other types of texts. First, sentences are parsed using a statistical parser. Then, for every main verb on the parse tree, a minimal clause structure is built. The initial clauses are refined and null elements on the parse tree are filled, by using verb subcategorization and verb semantics. Then, the semantic roles, adjuncts and verb meaning in each clause are computed. The algorithms have been tested on the fourth release of Ontonotes and Wikipedia.

Keywords: Semantics,Encyclopedic Texts, Natural Language Processing

1. Introduction

We describe algorithms for constructing a semantic representation for sentences from encyclopedic texts and other unannotated corpora. A semantic interpreter assigns meaning to the constituents in the sentence and determines semantic roles and adjuncts (temporal, spatial, etc.). Semantic interpreters rely on parsers to determine the syntactic structure of the sentence. Statistical parsers, which are becoming predominant, are very robust, but provide an impoverished parse tree in which long distance dependency and missing elements on the tree are not indicated, let alone filled with their referents. Linguists refer to these missing elements, or null elements, as empty categories. The recovery of these missing elements is a necessary condition for extracting semantic information from parse trees. Empty cat-

egories may result from diverse linguistic phenomena, i.e., relative clauses, missing complements, etc. A semantic interpreter requires that the empty nodes be recognized, and some of them filled with their antecedents.

Even if all empty nodes on the parse tree are filled, the problems of deciding which constituents are arguments and which ones are adjuncts remain. On the parse trees, produced by Charniak's parser (Charniak, 2000), for the sentences below all nodes are filled.

```
(1) She knew when John left.

 (S1 (S (NP (PRP She)) (VP (VBD knew) (SBAR (WHADVP (WRB when))
                     (S (NP  (NNP John)) (VP (VBD left)))))))

(2) She ate when John left.

 (S1 (S (NP (PRP She)) (VP (VBD ate) (SBAR (WHADVP (WRB when))
                    (S (NP (NNP John)) (VP (VBD left)))))))
```

However, in (1) the SBAR, subordinate clause (what she knew), is an argument, but in (2) the SBAR (subordinate clause) is a temporal adjunct.

The algorithms explained in this paper have four steps, each step receiving as input the output of the previous one. In step1, sentences are parsed using Charniak's parser (Charniak, 2000). In step2, all the clauses for each non-auxiliary verb on the parse tree are produced (section 3) by the MCR algorithm (minimal clausal reconstruction). In step3, clausal complements are recognized as arguments or adjuncts (section 4), and null subjects are filled (sections 5 and 6). Finally, in step4 the semantic interpreter (henceforth, SI) is activated to determine verb meaning, or verb predicate, semantic roles and adjuncts. The next section explains related research, and the last three sections in the paper deal with testing and conclusions. Figure 1 describes the final output of the semantic interpreter for the sentence "Achilles urged the Greeks to sail home." In my home page [1] under the title of this paper one can find examples of semantic interpretation for over 5,500.00 in two files "Verbs Tested" and "Sentences from Wikipedia, which give a much better idea of the capability of the system. These files, an integral part of this

[1] www.cs.ucf.edu/home/~gomez

```
 (g925 (subj ((noun achilles)) ((mythical_being1 achilles1)) (agent))
   (verb urged ((main-verb urge urged)) urge-someone (urge1)
               supported by  3 srs)
         obj ((dt the) (pn greeks)) ((european1 greek2)) (recipient))
           (obj2 (gensym g926 infinitive) (theme)))

urge-someone > advise > trans-infor > communicate > interact > action

(g926 (subj ((dt the) (pn greeks)) ((european1 greek2)) (agent))
 (verb sail ((main-verb sail sail)) sail (sail1) supported by 2 srs)
 (obj ((noun home)) ((residence1 home1)) (to-loc)))

sail > travel-by-boat > change-location > action
```

Figure 1: Output for the clauses of "urge" and "sail" in the sentence "Achilles urged the Greeks to sail home"

paper, contain the sentences, the output of the parser, the MCR algorithm (minimal clausal reconstruction) and the semantic interpreter for each clause in the sentence. Note that in the example in Figure 1, the parser has identified "home" as NP, although it is an adverb. However, the SI (semantic interpreter) has recognized it correctly as a *to-loc*, or *destaination*, role. The output of the semantic interpreter identifies the verb meaning, or verb predicate, the semantic roles, *agent, theme, recipient, to-loc* for these clauses. It also produces the super-predicates (indicated by >) for the verb predicate in each sentence.

2. Related Work

Most of the work to resolve empty categories on parse trees has been based on supervised learning algorithms using annotated corpora. A notable exception is Campbell's work (Campbell, 2004) which uses principles of Government and Binding (Chomsky, 1981) to detect and recover empty categories. His system outperformed previously learning-based algorithms on the detection and recovery of empty categories. Campbell's algorithm does not use lexical information, nor verb semantics. The algorithms explained in this paper are also rule-based algorithms that use general linguistic principles, but also verb lexical subcategorization and verb semantics, and they

go beyond Campbell's algorithms in that they also identify verb arguments, verb meaning and adjuncts.

The machine learning approaches to role labeling rely on syntactic features on the parse trees to determine the roles and label them (Gildea and Jurafsky, 2002). These algorithms use PROPBANK (Kingsbury et al., 2002), a resource annotated with verbs arguments but not with semantic roles and verb senses. The verbs annotated have low polysemy. Most roles are *arg0* (which corresponds to *agent*) and *arg1* (which corresponds to *theme*). Most times these two roles correspond to subject and object, and in many cases can be determined by using only syntactic features on the parse trees, including path features. A problem with path features is that the paths on the parse tree may vary greatly for sentences with very similar syntactic structures (Vickrey and Koller, 2008). More importantly, the semantics of nouns and verbs is critical to determine semantic roles and verb meaning. Consider the sentence "Joshua is bitten by a tarantula and is driven into a frenzy/hotel" in which the meaning of "drive" and the semantic role of the PP "into" vary depending just on the head noun of the PP ("frenzy" or "hotel.")

FRAMENET (Fillmore et al., 2003) is a corpus annotated with semantic roles that bears strong relation to this work, but a major difference is that FRAMENET frames are not linked to selectional preferences and grammatical relations. SHALMANESER (Erk and Pado, 2006) is a software tool that uses FRAMENET semantics, parsing and other NLP tools to assign semantic roles The authors describe it as "a shallow semantic parsing tool .." that "assigns semantic classes (senses) to words, and it assigns semantic roles. Both sense and role assignment are modeled as supervised learning tasks." We tested SHALMANESER on sentences from *The World Book Encyclopedia (@ 2011 World Book, Inc.)*, an encyclopedia which is very well edited. Our test showed that SHALMANESER failed to fill empty nodes and to assign semantic roles, and verb predicates for many of the clauses in these sentences.

Recently Ontonotes (Weischedel et al., 2011) offers an analysis of verb senses, its matching to WordNet (Fellbaum, 1998) verb senses, and an annotation of actual uses of the verbs. Ontonotes is a very valuable resource not only because of the annotation but by providing a classification of verb senses into groups. A critique that could be leveled is that some of the Ontonotes group senses for verbs with high polysemy are too coarse, coalescing metaphoric and literal senses of the verbs, which are clearly distinct.

The view of this research is that verb predicate, or verb meaning, and

grammatical relations are interconnected (Pinker, 1989). We are using the term "verb predicate" in the sense of Grimshaw (Grimshaw, 1990). Each verb predicate corresponds to a verb sense. For instance, two of the verb predicates for the verb "drive" are *to operate-a-vehicle* ("She drove the bus") (sense 1 in Wordnet) and *to compel-somebody* ("He was driven by his ambition.") sense 5 in WordNet. Each verb predicate has its own semantic roles which may be realized by different selectional preferences and grammatical relations. Because of the strong connection of verb meaning to syntax and semantics, this research has been strongly influenced by the work on argument structure (Chomsky, 1981), (Grimshaw, 1990), (Pritchett, 1992), (Jackendoff, 1990), and also by minimal commitment models (Marcus, 1987; Gorrell, 1991; Weinberg, 1993).

In Gomez (Gomez, 2001, 2004), we have shown how to define verb predicates for WordNet verb classes (Fellbaum, 1998), linked to grammatical relations and selectional preferences. The verb predicates have been mapped to WordNet verb classes (Fellbaum, 1998), and the selectional preferences to WordNet noun ontology (Miller, 1998). In principle, the definition of a verb predicate for a given WordNet class applies to all verbs under that class, which, in some cases, may contain about 2000 verbs. However, the verb predicate for a verb class fails to apply to many verbs under that class, because these verbs may realize their semantic roles by different selectional preferences and grammatical categories. As a result, those verbs that do not fit into a WordNet class need their own definitions, though they may inherit some of its roles from its super-class. See (Gomez, 2004) for a detailed description of these aspects. There has been work on automatically acquiring selectional preferences for verbs and adjectives (Resnik, 1996; McCarthy and Carroll, 2003; Erk, 2007; Tanner and Gomez, 2010). But, these attempts have achieved limited success, and most of this work has focused on subject verb, and verb object pairs without dealing with prepositions. Moreover, the acquired selectional preferences work for verbs with very low polysemy, such as "eat" or "drink" etc. But, our research concentrates on getting selectional preferences that will enable the determination of verb meaning, i.e., verb predicate and semantic roles for any verb and its roles whether those roles are realized by subject, objects, prepositions, or clauses.

In addition, we have described an algorithm that determines verb meaning and semantic roles, using the verb predicates. For every non-auxiliary verb in a sentence, a list of verb predicates for that verb is provided. The goals of the algorithm are to select one verb predicate from that list, thus determining the

sense of the verb, and to identify its semantic roles and adjuncts. All these tasks are simultaneously achieved. For each grammatical relation (GR) in the clause and for every verb predicate in the list of predicates, the algorithm checks if the predicate explains the GR. A predicate *explains* a GR if there is a semantic role in the predicate realized by the grammatical relation and the selectional preferences of the semantic role subsume the ontological category of the head noun of the grammatical relation. This process is repeated for each GR in the clause and each predicate in the list of predicates for the verb of the clause. Then, the predicate that explains the most GRs is selected as the meaning of the verb. The semantic roles of the predicate have been identified as a result of this process. Every constituent that has not been mapped to a semantic role must be an adjunct or an NP modifier. The entries for adjuncts are stored in the root nodes *action* and *description* (for stative verbs) and are inherited by all predicates in those categories. Adjuncts are identified after the predicate of the verb has been determined because adjuncts are not part of the argument structure of the predicate (Grimshaw, 1990). In Figure 2 we have included the definitions of two predicates for the verb "leave," out of a total of eleven predicates dealing with the 16 senses of "leave" in WordNet.

The predicates *leave-a-place* is for one of the senses of the verb "leave." Its super-predicate is *change-location*. This predicate captures the abstract event of a human or an animal leaving a place by changing location. The wn-map slot indicates its corresponding WordNet senses. Not all semantic roles are listed because some of them are inherited from its super-predicate. Each semantic role contains the selectional preferences, which are WordNet ontological categories, and the grammatical relations that realize them. Both the selectional preferences and the grammatical relations need to be true for the semantic role to be filled. For instance, the *agent* of *leave-a-place* is *human-agent* (which includes both humans and organizations) or *animal1* (animals). This role is realized by the grammatical subject in the sentence. The minus sign "-" preceding the categories *human-agent* and *animal1* means that those ontological categories are excluded from the role *to-loc*. The predicate *leave-give* is for the sense of "leave" when it means "bequeath", e.g. *He left a fortune to his children.* The syntactic entry "subj-if-obj" is realized by the subject of the sentence if the sentence also contains an obj (first post-verbal NP). The entry *obj-if-obj2* in the role *to-poss* is realized by the object of the sentence if the sentence contains also an obj2 (second post-verbal NP). The *theme*, the thing being transferred, can be realized by an obj, or an obj2.

```
[leave-a-place
     (is-a(change-location))
     (wn-map(leave1)(leave5))
     (agent(human-agent animal1) (subj))
     (from-loc(location) (obj))
     (toward-loc( location  physical-thing)((prep  towards toward)))
     (to-loc(location -human-agent -animal1  physical-thing) ((prep for)))
     (instrument(vehicle )((prep with on in by))
               (animal1)  ((prep on)))
     (to-do-an-activity(action event )((prep for)))]


 [leave-give
     (is-a(give))
     (wn-map(leave12)(leave15))
     (agent(human-agent)(subj-if-obj))
     (to-poss (human-agent animal1 ) (obj-if-obj2)
              (human-agent animal1)  ((prep  to)))
    (theme(possession thing) (obj obj2))
    (thematic-rule(require(to-poss)))]
```

Figure 2: Definitions for the predicates "leave-a-place" (change-location) and "leave-give" (transfer-of-possession) for the verb "leave"

The entry "thematic-rule" means that the predicate requires the role *to-poss*. If that role is not filled, that predicate will not be chosen.

Consider the sentence "The farmer left the orchard." The *agent* of *leave-a-place* and *leave-give*, as well as the *agent* of other predicates for "leave" not listed, are realized by the subject, "the farmer." A farmer is a human and is the subject of the sentence. The role *to-loc* of *leave-a-place* will be realized by "the orchard." "Orchard" is a location in WordNet and is the object of the sentence. The *theme* of *leave-give* is also realized by "the orchard." But, because "leave-give" requires a *to-poss* role which has not been filled, the predicate "leave-a-place" is chosen. But, suppose that the sentence is "The farmer left the orchard to her daughter." In this case, the role *to-poss* in *leave-give* will be realized by *to her daughter*, while there are no roles in *leave-a-place* that matches *to her daughter*. Because *leave-give* explains more semantic roles, it is chosen as the meaning of the verb. Thus, verb meaning becomes a side effect of identifying semantic roles. Sometimes, several predicates may explain the same number of semantic roles. In those cases, the algorithm chooses the first predicate defined and lists all those predicates tied to it. For instance, in the sentence "Several students left the university" the algorithm will select two predicates: *leave-a-place* and *abandon-an-activity-or-a-position*. The strong relation between grammatical relations and selectional preferences is illustrated in sentences such as "The farmer left the tractor in the orchard," meaning *leave-behind*, or "The farmer left the orchard in ruins," meaning *leave-something-in-a-state*, in which the grammatical relations and the ontological categories of the head nouns of the NPs determine the semantic role, and, as a result, the verb meaning, or verb predicate.

This algorithm depends on the parser to determine the clauses and grammatical relations in the sentence. Thus, the parser will indicate that the SBAR (subordinate clause) in "She knew when John left" is a complement of "knew", but in "She ate when John left" the SBAR clause is a temporal adjunct. In addition, the algorithm requires from the parser to fill the null elements on the parse trees. These are two major requirements that statistical parsers do not meet. The parser used by the semantic interpreter algorithms described in (Gomez, 2004; Gomez and Segami, 2007) can determine grammatical relations and resolve long distance dependency with about 68% correctness in *The World Book Encyclopedia (@ 2011 World Book, Inc.)*, if the length of the sentences is about 17 words. However, for longer sentences the correctness is much lower. Statistical parsers perform better and are

more robust that the parser used in (Gomez and Segami, 2007). Moreover, improvements in statistical parsers are immediately translated into improvements in the system described here. The MCR algorithm (minimal clause construction algorithm) explained in the next section is based on our earlier parser, and can be viewed as an adaptation of some of the ideas in that parser to the output of statistical parsers.

3. Minimal Clausal Reconstruction (MCR) from Constituent-Based Parsed Trees

We approach the analysis of parse trees from the point of view that a sentence contains one or more clauses, and each clause has its main verb, arguments and/or adjuncts. The approach bears certain similarity to Winograd's (Winograd, 1983) analysis of systemic grammars for parsing. The task of the MCR algorithm is to identify each clause in the sentence, and its constituents. The input to the MCR algorithm is the output of the statistical parser (Charniak, 2000). The MCR algorithm reconstructs the clauses of the sentence from the parse tree in a minimal way (not related to minimalism), because the MCR does not decide which constituents of the clause are arguments, and which ones are adjuncts, and build a list of possible subjects for some clauses when the parse tree has an empty node for the subject position. However, the MCR does resolve long distance dependency resulting from relative clauses. By constructing a clause structure for every main verb on the parse tree, the MCR discovers syntactical similarities between clauses and sentences, avoiding the problems discussed in (Vickrey and Koller, 2008).

The major tasks are: 1) to recover the verb sequence, indicating whether the sentence is passive or active, 2) to recover all post-verbal constituents of the VP (all those constituents dominated by the VP), 3) fill subject gaps, 4) break up attachment of post-verbal PPs to NPs, and 5) resolve long distance dependency resulting from relative clauses. The post-verbal constituents are enumerated from left to right as obj, obj2, obji. In a parse tree, these are all constituents dominated by the VP node. These constituents can be NPs, VPs, SBARs or Ss. Consider the parse tree for the sentence "She told the children that Mary left."

```
(S1 (S (NP (PRP She)) (VP (VBD told) (NP (DT the) (NNS children))
        (SBAR (IN that) (S (NP (NNP Mary)) (VP (VBD left)))))))
```

which the MCR algorithm transforms into:

```
(g864 (subj((pron she)) verb ((main-verb tell told) (tense vbd))
         obj ((dt the) (noun children))
          obj2 ((gensym g865 that-clause that)))

g865 (subj ((pn mary)) verb ((main-verb leave left) (tense vbd))
         type  (that-clause that) parent-verb (tell g864)))
```

We are using the Lisp gensyms (g864, g865), "G" concatenated with a string of numbers, to name the clauses produced by the MCR. The output for the sentence is an association list in which the name of the clause, the gensym, precedes it. The content of obj2 in the first structure contains a gensym, g865, which is used as a pointer to the actual structure for g865. The MCR is not saying that the clause for "tell" has two post-verbal complements/arguments as it happens in this case, but rather that it has two constituents. It is up to the SI (semantic interpreter) to decide if they are actually arguments or adjuncts.

The other main task of the MCR is to recover subjects but in a minimal way. If the subject of a clause is missing on the parse tree, the MCR algorithm chooses the subj and obj of the main clause *in that order* as possible subjects, by opening a Subj entry in the structure and inserting them in it. For the sentence "She bought a book of history to learn the truth," the parse tree is:

```
(S1 (S (NP (PRP She)) (VP (VBD bought) (NP (NP (DT a) (NN book))
        (PP (IN of) (NP (NN history))))
      (S (VP (TO to) (VP (VB learn) (NP (DT the) (NN truth))))))))
```

which the MCR transforms into:

```
(g880
 (subj ((pron she)) verb ((main-verb buy bought) (tense vbd))
    obj ((dt a) (noun book)) prep (of ((noun history)))
                obj2 ((gensym g881 infinitive)))

 g881
 (subj (((pron she)) (((dt a) (noun book)) prep (of ((noun history)))))
   verb ((main-verb learn learn) (tense vb)) type (infinitive)
   parent-verb(buy g880)  obj ((dt the) (noun truth))))
```

The obj2 in clause G880 is not an argument of "bought," and "she" is the subject of "learn." The next step will decide about this. For "She

promised/asked Mary to read a book," the MCR builds the same two possible subjects for "read." Again, the next step will choose the subject of "read" depending on the subcategorization for "promise" and "ask." For VP infinitives dominated by NPs (e.g, "His ability to understand Physics impressed everybody.") the MCR builds *np-dominated-infinitive*. The SI will have its say in deciding if, in fact, a noun in the NP subcategorizes the infinitival clause.

If the main clause follows a subordinate clause without a subject as it happens in reduced adverbial clauses in front of sentences (e.g., "While visiting Iowa, she bought a farm."), the subject assigned to the subordinate clause is the subject of the main clause. The MCR builds the verb sequence, identifying passive clauses. If the clause is passive, it does not change the subj to obj. But if the verb has a post-verbal constituent, the MCR names it obj2. Hence, for the sentence "He was told the truth," the MCR builds

```
(subj ((pron he)) verb ((aux (was)) (main-verb tell told)
 (tense vbn) (voice passive)) obj2  ((dt the) (noun truth)))
```

In recovering subjects of clauses dominated by passive clauses, the MCR builds *unknown-agent* as the subject of the embedded clauses. Thus, for the sentence "He was told to get a loan," the possible subjects for "get" are *unknown-agent* and "he," in that order. For "The houses were bought to be sold," the possible subjects of "sold" are *unknown-agent* and "houses" in that order. If the passive clause has also a [by NP] constituent, the MCR builds it as a possible subject for the embedded clause. As mentioned, the SI (semantic interpreter) will choose the correct subject among the possible subjects passed by the MCR.

Another major aspect of the MCR algorithm is that it breaks up post-verbal prepositional phrases attached to NPs by the parser, except under apposition and coordination. Thus, for the sentences "He rid the city of rats," "She made a statue of marble," or "Farmers enrich the soil with fertilizers," the MCR algorithm will detach "of rats," "of marble," and "with fertilizers" if they were attached to the NP by the parser. We want the argument structure of the verb predicate to have the first say in attaching post-verbal prepositional phrases. This is in agreement with the ideas expressed in (Pritchett, 1992) and with minimal commitment models (Marcus, 1987; Gorrell, 1991; Weinberg, 1993). If the SI attaches the PPs to the verb *strongly*, the parser's attachment is overridden. However, if the SI attaches

the PP *weakly*, the parser's attachment is preferred. Obviously, all verb arguments realized by prepositions are *strongly* attached to the verb, while adjuncts are attached *weakly*. Adjective phrases (ADJP) are mapped into PRED, which stands for adjectival predicate.

4. Refining the output of the MCR

The major tasks in the algorithms that refine the output of the MCR are explained next. For each clause produced by the MCR, starting with the first clause outputted, do: 1) recognize clausal complements, 2) select from possible subjects, 3) verify if the sentence is actually passive/active by using verb subcategorization and the clause structures built for the sentence, 4) activate the SI (semantic interpreter) with that clause.

When the nodes dominated by the VP are clauses, the output of the MCR needs to be cleaned using the verb subcategorization information, prior to activating the SI. The critical entries are the nodes obj and obj2, which, when used in the verb predicate entries, stand for complements. However, in the output produced by the MCR these entries may stand for subordinate clauses not subcategorized by the main verb. Possible NP complements are left to the SI (semantic interpreter) to decide if they are actually arguments, or adjuncts.

Let us use the variable OBJ-CLAUSE to refer to any obj, obj2, ... obj$i$ whose content is a clause, a gensym. The algorithm for recognizing clausal complements, described in Figure 3, is activated for each OBJ-CLAUSE in the structure from left to right. If the algorithm determines that the OBJ-CLAUSE is a CP (complement phrase), the OBJ-CLAUSE is left in the structure to be processed by the SI (semantic interpreter), else the OBJ-CLAUSE is erased from the structure, and the index of any obj$i$ still in the structure is decreased by 1. Thus, obj$i$ is replaced with obj$i$*-1*. In "She told the children to study," obj is "the children" and obj2 is "to study," a clause. Thus the algorithm in Figure 3 needs to decide if obj2 is a complement of "tell," in which case it will be left in the structure as in this case, to be processed by the SI (semantic interpreter). In "He promised the children a book if they studied," the output of the MCR for "promise" is:

```
(g864
 (subj ((pron she)) verb ((main-verb promise promised) (tense vbd))
     obj ((dt the) (noun children)) obj2 ((dt a) (noun book))
               obj3 ((gensym g865 if-clause if))))
```

rcp1. If the OBJ-CLAUSE is a VP infinitive and the main verb subcategorizes the infinitive return T. Else assume the infinitive is a purpose infinitive and remember that fact.
rcp2. If the OBJ-CLAUSE is a VP VBG (e.g., "He likes eating apples.") and the main verb subcategorizes a VP VBG, return T.
rcp3. If the OBJ-CLAUSE is a that-clause, or a s-c (empty sbar) clause and the main verb subcategorizes a that-clause or the root-form of the verb, return T.
rcp4. If the clause type in OBJ-CLAUSE is a WHADVP and the main verb subcategorizes the conjunction return T.
rcp5. If the clause type in OBJ-CLAUSE is WHNP and the main verb subcategorizes a WHNP, return T.
rcp6. Else return NIL.

Figure 3: Algorithm for Recognizing Clausal Complements (RCP)

The algorithm needs to decide if obj3, an OBJ-CLAUSE, is a complement of "promise." It is not a complement because it is not subcategorized by "promise," so obj3 is erased from the structure. In our testing, few cases popped up in which a verb subcategorization is not in our database.

5. Algorithm for Choosing between Subjects (ACS)

When the MCR builds more than one subject for a clause, the algorithm (ACS) chooses between subjects based mainly on the verb subcategorization, but also on the semantic interpretation of the clause already interpreted. The general cases are given in Figure 4. Remember that possible subjects are entered in the following order: first the subject and then the obj of the main clause. Thus, consider the sentence "She bought a book of history to learn the truth," discussed in section 2. The potential subjects of "learn" are "she" and "book." The *Algorithm for Recognizing Clausal Complement* (Figure 3) would recognize "to learn" as a purpose infinitive on step rcp1, because "buy" does not subcategorize a VP infinitive. The ACS algorithm in Figure 4 selects "she" as the subject of "learn" in step1.2. For the sentence, "The houses were bought to be sold" the potential subjects of "sold" are "unknown-agent" and "the houses" in that order, and the ACS algorithm will select "houses" as the subj (step1.1), because the main clause is passive. For "He was told to be fair" the subjects entered for "be" are "unknown-

agent" and "he" in that order. The VP infinitive "to be" is recognized as an argument not as a purpose infinitive by the *Algorithm for Recognizing Clausal Complement* because it is subcategorized by "tell" (rcp1). and this algorithm (ACS) in Figure 4 (step2.2) will select the second entry in the subject slot, namely"he," as the subject of "be."

In step3, the ACS algorithm deals with sentences such as "Huge ice blocks prevented him from going farther." Because "prevent" subcategorizes "from," the second entry in the subj slot,"him," is selected. In step3.1 the algorithm also checks if the preposition belongs to the verb predicate of the main clause. This is so because there are still some prepositions missing from some verb subcategorizations, but these prepositions are present in one of the verb predicates for that verb. For instance, we did not have "from" as subcategorized by the verb "proscribe." But, "proscribe" is mapped to the verb predicate *forbid*, and the *theme* of *forbid* (thing or event being forbidden) is realized by "from" followed by a VP VBG clause. By accessing this verb predicate we could choose correctly "some people" as the subject of "giving" in "Their codes of ethics may proscribe some people from giving secret briefings." Step 4 defaults to the first entry in the subject slot.

6. Analyzing the subject of the CP in those cases in which the MCR builds only one subject

Not all infinitival clauses fill their null subject with the subject or the object of their main clause. For instance, "arrange" and "plan" may subcategorize infinitives whose subjects are introduced by a PP [for NP]; e.g. "Alexius arranged for Maria to stay on the palace." Charniak's parser (Charniak, 2000) does quite well in these types of verbs, recognizing perfectly the object of the PP in the main clause as the subject of the CP (complement phrase). However, Charniak's parser does not handle well other types of verbs in which the subject of the CP is a PP headed by "with," "to," or other prepositions. For instance, "She pleaded with him to read the books," "They appealed to the governor to save the trees," and "He yelled at her to stop." The verb subcategorization is checked for each verb before choosing a subject, even if the MCR builds only one subject. In verbs of transfer of possession, the null subject of the infinitives depends on the semantic roles and verb predicate of the main clause. In these verbs, the *to-poss* (the entity receiving the object being transferred) of the main clause is the subject of the infinitival clause. Consider the sentences "The war gave Japan an opportunity to enlarge its

- step1. If the clause type is an infinitive, and it has been recognized as a purpose infinitive by the algorithm that recognizes clausal complements (RCP) then,
  - step1.1 If the parent clause is passive, select the second entry in the subject slot.
  - step1.2 Else return the first entry in the subject slot.
- step2. If the clause type is an infinitive and it has been recognized as a complement by the RCP algorithm then,
  - step2.1 If the verb of the parent clause subcategorizes the subject (e.g. "promise") return the first entry in the subject slot.
  - step2.2 If the verb of the parent clause subcategorizes the obj, return the second entry in the subject slot.
  - step2.3 Else return the first entry - this happens rarely when missing subcategorizations.
- step3. If the clause type is [Prep VP VBG] (e.g., "from entering," "for forcing"), then
  - 3.1 If the main verb subcategorizes the preposition or the preposition is found in the verb predicate of the parent clause return the second entry in the subject slot.
  - 3.2 Else return the first entry in the subject slot
- step4. Return the first entry in the subject slot.

Figure 4: Algorithm for Choosing between Subjects (ACS)

empire" and "Habeas corpus guarantees a person under arrest a chance to be heard in court" (*The World Book Encyclopedia (@ 2011 World Book, Inc.)*). In these sentences "Japan" is the subject of "gave" and "a person" the subject of "be heard." Compare those sentences to "Businessmen saw a chance to make money," in which the subject of "make" is "businessmen." Our system incorporates two general semantic rules (3 lines each) that select the *recipient* role of transfer of information verbs, and the *to-poss* role of transfer of possession verbs in the main clause as the subject of infinitival clauses, and also as the subject of PP-VBG clauses headed by "of," e.g. "Pundits gave Clinton little chance of defeating Bush." These two rules use the semantic roles and verb predicates already identified for the main clauses. This is very much in line with Jackendoff's observations (Jackendoff, 1987) that the null subject of purpose and relative infinitives are controlled by the thematic role hierarchy.

Besides this rule, there are other rules to recognize "it" when it stands for a clausal substitute (Baker, 1995), and the clause as a pseudocomplement. For instance, "It is improbable that Athens would have sent twenty vessels ..." in which "Athens would have sent twenty vessels" is the subject of "is" and "it" is acting as the substitute for that clause. Other sentences within this category are "It has been said that she saved the country," or "It was widely believed that the war would be short." There are also rules to handle the *easy* construction (Baker, 1995), e.g., "It is easy/difficult/possible to copy books." The MCR will take "it" as the subject of both "was" and "copy," which is wrong. A rule will assign *unknown-agent* as the subject of "copy" and "to copy the books" as subject of "is."

A set of rules handles nouns (*ability, chance, effort, attempt, etc.*) that subcategorize clauses,, e.g., "She has provided further evidence of the ability of chimpanzees to understand language"' in which the subject of "understand" is "chimpanzees." In "The demonstrators protested against the efforts of white officials there to deny most black citizens the chance to register and vote," the subject of "deny" is "white officials" and the subject of "register" and "vote" is "black citizens." In "Everett was impressed by Lincoln's logic and ability to say so much in so few words," the subject of "say" is "Lincoln." (These sentences are from (*The World Book Encyclopedia (@ 2011 World Book, Inc.)*). The parser outputs these infinitival clauses on the parse tree as dominated by the NP of the nouns "chance," "ability," etc. The output for "Mary has the ability to see in the dark" is:

```
(S1 (S (NP (NNP Mary))(VP (AUX has) (NP (DT the)(NN ability)
     (S (VP (TO to) (VP (VB see)
        (PP (IN in) (NP (DT the) (NN dark)))))) ))))
```

Note that the NP for "ability" dominates the infinitival clause (the parenthesis preceding the NP for "the ability" is closed after the VP infinitive "to see." Unfortunately, the parser fails to recognize some clauses as subcategorized by some nouns, while in other cases the parser incorrectly identifies a clause as subcategorized by the noun. The grammatical rules repair the parser when incorrectly identifies a clause as subcategorized by the noun. All the details of these grammatical rules to handle the subjects of clauses subcategorized by nouns will be published in a separate paper.

7. Testing the Algorithms

The testing cannot be automated because of the different data structures in Ontonotes (Weischedel et al., 2011) and in the MCR. Every clause has to be manually checked for correctness by looking into the Ontonotes files. We selected randomly a letter in Ontonotes release 3 (Ontonotes release 4 had not been published yet when the test took place.) with at least 50 verbs, and map all senses for those verbs to the verb predicates. The letter "m" was selected, which in the release 3 of Ontonotes contains 76 verbs. We have mapped all the Ontonotes verb senses except 5 verbs (monosemous and none occurring in the WSJ annotated corpus) to verb predicates. We checked the definitions on sentences from the *The World Book Encyclopedia (@ 2011 World Book, Inc.)* and the *BNC corpus (@ 2010 University of Oxford)*. As our definitions were ending Ontonotes release 4 was published. Then, all our testing has taken place against the annotated data in Ontonotes4. For each verb, our program produces the WordNet (WN) senses corresponding to the Ontonotes verb groups. If a verb has more senses in Ontonotes than in any version of WN, our program produces the Ontonotes verb group. For some verbs, our senses are more fine grained than those in Ontonotes. We kept our senses and judge the tested sentences correctly if our verb senses are all contained in the Ontonotes verb group selected for that verb.

We have tested each verb on at least 30 sentences. "Maximize" has been tested with only the 5 sentences in the WSJ files because we had already tested "minimize" with 31 sentences, and they have the same semantic and grammatical distribution. During the testing, there are many verbs in the

WSJ Ontonotes files which are not annotated with verb senses. We have selected only those sentences for which the Ontonotes annotators provide verb senses. If there were less than 30 sentences in the WSJ file, we randomly selected the remainder sentences from Wikipedia. Also, if only one verb sense was represented in the WSJ file, we selected additional Wikipedia sentences with several verb senses for the target verb. For "make" (18 senses), "move" (9 senses), and "mark" (6 senses), we selected 40 or more sentences from the WSJ files, and similar number from Wikipedia. In my home page (www.cs.ucf.edu/home/~gomez), the zipped folder "Verbs Tested," (immediately under the title of this paper) contains all the files for each verb, and their evaluation. There is a total of 74 files because "make,' "move," and "mark" have two files each. A total of 2505 sentences containing the target verbs have been graded by two graders. We parsed all the sentences including the ones in the Ontonotes with Charniak's parser (Charniak, 2000). Each verb file contains the sentences, their parses, the MCR output for the entire parse, and the verb predicate, semantic roles and adjuncts for the verb being tested, and its evaluation. Note that all clauses need to be interpreted in order to compute the roles, verb predicates and adjuncts. That is the only way on which the arguments for the target verb can be computed.

For adjuncts, we graded locative, temporal and event types, which are the most frequently occurring in the corpus tested. In the case of adjuncts, there are two problems: recognizing that they are adjuncts, and naming them as temporal, locative, beneficiary, etc. The first problem is very well defined except in the case of those roles that Pritchett (Pritchett, 1992) call quasi-arguments, e.g, *instrument*. The second problem is harder because there are many adjuncts which have not been properly categorized. Grouping many of them under the label *manner* is not satisfactory. In naming adjuncts, our system provides multiple labels. Thus, for the PP "for the population" in the sentence "An economic bonanza continued for the population" the adjunct is named as *beneficiary* or *purpose*. We have graded the adjunct as correct if the first name listed is the correct one.

The SI correctly identifies *purpose infinitives* in most cases but they have not been graded. The grammatical relation *subject* has been graded and is listed on the tables. All grammatical subjects have been graded. If a subject is the wrong subject, the semantic role for that grammatical relation is not graded as wrong or correct. If a PP is weakly attached by the algorithms to the verb, but the parser attaches it to a NP, the attachment of the parser is preferred.

-	CORRECT	MISSED	WRONG
pred	1742	2	64
roles	3243	92	68
adjuncts	247	20	16
subject	1717	0	91

Table 1: Values for predicates, roles, adjuncts and subject for verbs with 3 or less senses

We could have selected sentences in which the heads of the constituents are not pronouns, or numbers, since we have used a very simple anaphora resolution algorithm. However, we opted against that when "it" starts a sentence. We initialized pronouns to some ontological categories, and let the SI choose. For instance, these are the ontological categories for "it:" *(it (is-a (animal1) (organization) (location) (physical-thing) (abstraction) (thing)))*. These are the ontological categories for "one:" *(one (is-a (human) (social-group) (animal1) (physical-thing) (abstraction) (thing)))*. "It" is of especial interest when it starts a sentence, because we want to see if the algorithms recognize "it" as a clausal substitute. Proper nouns present similar problems, not only because there are few proper nouns in WN, but some of them do not have all its ontological categories. For instance, "Hittite" appears only as a language. If the proper noun is not in WN, the following ontological categories are assumed: *human, social-group, location, written-communication.* The ontological categories assumed for common nouns (NN/NNS) not in WN are: *human, social-group, animal1, physical-thing, abstraction, thing.* Because our semantic roles distinguish between "agent" (a human, a social group or animal), and *inanimate-cause* (a causal agent other than a human-agent or an animal), some arguments are labeled with those two roles, which correspond to the argument *arg0* in Ontonotes. Thus, in the sentence "The cranes destroyed the fields", "the cranes" is identified as *inanimate-cause* (sense selected crane1) or *agent* (crane2) (depending on the meaning of "crane," but in "The cranes ate the insects," "the cranes" is identified only as *agent*, selecting only the bird sense of "crane," crane2.

Table 1 provide the results for all those verbs tested containing 3 or less senses. The polysemy breakdown in this group is as follows: there are 12 monosemous verbs, 22 verbs with two senses, and 21 verbs with 3 senses. The entry "pred" stands for verb predicate, or verb sense. The verb sense has been identified correctly in 1742 verb occurrences out of 1808. The entries for "adjuncts" and "roles" have similar readings. The subject of the clause

-	P	R	F1
pred	96.46	99.89	98.14
roles	97.95	97.24	97.59
adjuncts	93.92	92.51	93.21
subject	94.97	100.00	97.42

Table 2: Precision, Recall and F1 values for Table 1

-	CORRECT	MISSED	WRONG
pred	645	0	52
roles	1227	38	37
adjuncts	162	7	9
subject	651	0	46

Table 3: Values for predicates, roles, adjuncts and subject for verbs with 4 or more senses

has been identified correctly in 1717 out of 1808 sentences for these verbs. Table 2 provides the precision, recall and F1 values for table 1.

Table 3 and 4 lists similar information for those verbs tested with 4 or more senses. In this group, the breakdown of polysemy is: 9 verbs with 4 senses, one verb with 5 senses, 3 with 6 senses, one with 8 senses, one with 9 senses, and one with 18 senses. If one compares table 2 and table 4, the entry for the verb predicates (verb senses) has the higher difference because of the greatest polysemy of verbs on table 3. But, there is not a major difference between the other entries on the table. Roles can be determined with great degree of accuracy for all verbs, but determining verb predicates is a harder problem. We have indicated as correct only the first verb sense, or predicate, selected by the system, but there are cases in which the second choice selected is the correct one. A noun disambiguation algorithm applied to the output of the the semantic interpreter can help to choose between tied predicates.

8. Additional Testing in Wikipedia

The goal of this research is to construct semantic structures for unannotated corpora with minimal human intervention, or without it, in order to facilitate question-answering, information retrieval and knowledge extraction from these texts. We randomly selected 3018 sentences from Wikipedia, containing the word "war(s)" in any position. In this test, we are not targeting

-	P	R	F1
pred	92.54	100.00	96.13
roles	97.07	97.00	97.03
adjuncts	94.74	95.86	95.29
subject	93.40	100.00	96.59

Table 4: Precision, Recall and F1 values for Table 3

any specific verbs. The maximum length of the sentences is 35 words. These files can be found in my home page (www.cs.ucf.edu/~gomez), in the zipped folder "Sentences from Wikipedia" (immediately after the title of this paper). The total number of clauses in these sentences is 6,793.00, containing 1776 distinct verbs. Phrasal verbs, e.g. break off, were considered distinct verbs, e.g., "break off" different from "break." The output of the parser was analyzed and if the parse contained fatal errors (a noun taken as a verb, or a verb as noun) were excluded since these sentences will produce irrecoverable interpretation errors. However, if the parse was mostly correct containing only errors in some clauses, the parse was interpreted and its output added to the file. Unfortunately, Charniak's parser as it is now makes about 25% of errors in the Wikipedia sentences.

Hence, in examining the files it is important to keep in mind that some semantic interpretation errors are caused by the parser errors. For instance, in the sentence "Germany's main concern in the Baltic sea was to protect the routes which supplied its war industry with vital iron ore imported from Sweden," the parser fails to recognize "imported," as a VP, consequently the SI fails to produce an interpretation for that clause, but the other clauses in the sentence are correctly interpreted. In addition to excluding sentences with bad parses, if a verb sense did not fit within one verb class, we define verb predicates for it.

We have evaluated 2 of the files, firstwar.i with 420 sentences, and secwar.i with 512 sentences. This is approximately 1/3 of all sentences. The evaluation has checked only if the verb predicate and the subject of the clauses are correct. Because in our approach the verb predicate and the semantic roles are interconnected, 90% of the clauses in which the verb predicate is correct all the semantic roles are also correct. This requires less time to evaluate and automatically generate corpora annotated with verb meaning and semantic roles. Note that a manual effort has gone in eliminating sentences with incorrect parses. The first file with 920 clauses has only 27 verb predicates and

21 subjects incorrectly identified, while the second file with 1166 clauses has 50 verb predicates and 29 subjects incorrect.

We are looking into producing semantic interpretation for a very large number of sentences from *The World Book Encyclopedia (@ 2011 World Book, Inc.)*. The sentences in this Encyclopedia are very well edited and are shorter than the ones in Wikipedia, and as a result the Charniak's parser makes less errors. This will facilitate the construction of semantic structures for this encyclopedia because fewer sentences will need to be checked for parser errors. Our initial tests on 2,224 sentences from this encyclopedia indicate that semantic structures like the ones explained in this paper can be constructed correctly for most of its sentences.

9. Conclusions

We have presented algorithms that determine verb predicates, semantic roles and adjuncts from statistical parsers. The main steps in the algorithms are the construction of a clause structure for every non-auxiliary verb on the parse tree, and the refinement of the clause structure by using verb subcategorization. The final computation of verb meaning and verb arguments is done by using selectional preferences in the verb predicates, which are linked to the Wordnet noun ontology. The algorithms have been tested on sentences from Ontonotes and Wikipedia. The integration of statistical parsers in the overall system architecture makes it possible to determine verb meaning and semantic roles for unrestricted texts. As statistical parsers improve, so will the overall performance of our system. The grammatical and semantic knowledge in the system can be easily refined and extended as it is run on different texts.

References

Baker, C.L., 1995. English Syntax. The MIT Press, Cambridge, MA.

Campbell, R., 2004. Using linguistic principles to recover empty categories, in: Proc. of the 2004th Annual Meeting of the ACL, pp. 645–652.

Charniak, E., 2000. A maximum-entropy-inspired parser, in: Proceedings of the North American Chapter of the Association for Computational Linguistics, pp. 132–139.

Chomsky, N., 1981. Lectures on Government and Binding. Foris, Dordrecht, The Netherlands.

Erk, K., 2007. A simple, similarity-based model for selectional preferences, in: Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, Prague, Czech Republic. pp. 216–223.

Erk, K., Pado, S., 2006. Shalmaneser - a flexible toolbox for semantic role assignment, in: Proceedings of LREC 2006, Genoa, Italy. pp. –.

Fellbaum, C., 1998. A semantic network of english verbs, in: Fellbaum, C. (Ed.), WordNet: An electronic Lexical Database and some of its applications. MIT Press, 1998, Cambridge, Mass, pp. 69–104.

Fillmore, C.J., Johnson, R., C., Petruck, M., 2003. Background to framenet. International Journal of Lexicography 16, 235–250.

Gildea, D., Jurafsky, D., 2002. Automatic labeling of semantic roles. Computational Linguistics 28, 245–288.

Gomez, F., 2001. An algorithm for aspects of semantic interpretation using an enhanced wordnet, in: Proccedings of the 2nd North American Meeting of the North American Association for Computational Linguistics, NAACL-2001, pp. 87–94.

Gomez, F., 2004. Building verb predicates: A computational view, in: Proccedings of the 42nd Meeting of the Association for Computational Linguistics, ACL-04, Barcelona, Spain. pp. 351–358.

Gomez, F., Segami, C., 2007. Semantic interpretation and knowledge extraction. Knowledge-Based Systems 20, 51–60.

Gorrell, P., 1991. Subcategorization and sentence processing, in: Berwick, R., Abney, S., Tenny, C. (Eds.), Principle-Based Parsing: Computation and Psycholinguistics. Kluwer Academic, Dordrecht, The Netherlands.

Grimshaw, J., 1990. Argument Structure. MIT Press, Cambridge, Mass.

Jackendoff, R., 1987. The status of thematic relations in linguistic theory. Linguistic Inquiry 18, 369–411.

Jackendoff, R., 1990. Semantic Structures. MIT Press, Cambridge, Mass.

Kingsbury, P., Palmer, M., Marcus, M., 2002. Adding semantic annotation to the penn treebank, in: Proc. of the Human Language Technology Conference, San Diego, CA. pp. –.

Marcus, M., 1987. Deterministic parsing and description theory, in: Whitelock, P., Wood, M., Somers, H., Johnson, R., Bennett, P. (Eds.), Linguistic Theory and Computer Applications. Academic Press.

McCarthy, D., Carroll, J., 2003. Disambiguating nouns, verbs, and adjectives using automatically acquired selectional preferences. Computational Linguistics 29, 639–654.

Miller, G., 1998. Nouns in wordnet, in: Fellbaum, C. (Ed.), WordNet: An electronic Lexical Database and some of its applications. MIT Press, 1998, Cambridge, Mass, pp. 23–46.

Pinker, S., 1989. Learnability and Cognition. MIT Press, Cambridge, Mass.

Pritchett, B.L., 1992. Grammatical Competence and Parsing Performance. The University of Chicago Press, Chicago,Illinois.

Resnik, P., 1996. Selectional constraints: An information-theoretic model and its computational realization. Cognition 61, 127–159.

Tanner, J., Gomez, F., 2010. Extracting ontological selectional preferences for non-pertainym adjectives from the google corpus., in: Proceedings of The Twenty-Fourth AAAI Conference on Artificial Intelligence, pp. 1033–1038.

Vickrey, D., Koller, D., 2008. Sentence simplification for semantic role labeling, in: Proceedings of ACL-08, Columbus, Ohio. pp. 344–352.

Weinberg, A., 1993. Parameters in the theory of sentence processing: Minimal commitment theory goes east. Journal of Psycholinguistic Research 22, 339–364.

Weischedel, R., Palmer, M., Marcus, M., Hovy, E., Pradhan, S., Ramshaw, L., Xue, N., Taylor, A., Kaufman, J., Franchini, M., et alia, 2011. Ontonotes release 4.0. Linguistic Data Consortium, Philadelphia, Penn.

Winograd, T., 1983. Language as a Cognitive Process: Volume 1, Syntax. Addison-Wesley Publishing Company, Reading, Mass.