(ALMOST) AUTOMATIC SEMANTIC FEATURE EXTRACTION FROM TECHNICAL TEXT

Rajeev Agarwal[*]
rajeev@cs.msstate.edu

Department of Computer Science
Mississippi State University
Mississippi State, MS 39762.

ABSTRACT

Acquisition of semantic information is necessary for proper understanding of natural language text. Such information is often domain-specific in nature and must be acquired from the domain. This causes a problem whenever a natural language processing (NLP) system is moved from one domain to another. The portability of an NLP system can be improved if these semantic features can be acquired with limited human intervention. This paper proposes an approach towards (almost) automatic semantic feature extraction.

1. INTRODUCTION

Acquisition of semantic information is necessary for proper understanding of natural language text. Such information is often domain-specific in nature and must be acquired from the domain. When an NLP system is moved from one domain to another, usually a substantial amount of time has to be spent in tailoring the system to the new domain. Most of this time is spent on acquiring the semantic features specific to that domain. It is important to automate the process of acquisition of semantic information as much as possible, and facilitate whatever human intervention is absolutely necessary. Portability of NLP systems has been of concern to researchers for some time [8, 5, 11, 9]. This paper proposes an approach to obtain the domain-dependent semantic features of any given domain in a domain-independent manner.

The next section will describe an existing NLP system (KUDZU) which has been developed at Mississippi State University. Section 3 will then present the motivation behind the automatic acquisition of the semantic features of a domain, and a brief outline of the methodology proposed to do it. Section 4 will describe the different steps in this methodology in detail. Section 5 will focus on the applications of the semantic features. Section 6 compares the proposed approach to similar research efforts. The last section presents some final comments.

2. THE EXISTING KUDZU SYSTEM

The research described in this paper is part of a larger ongoing project called the KUDZU (Knowledge Under Development from Zero Understanding) project. This project is aimed at exploring the automation of extraction of information from technical texts. The KUDZU system has two primary components — an NLP component, and a Knowledge Analysis (KA) component. This section describes this system in order to facilitate understanding of the approach described in this paper.

The NLP component consists of a tagger, a semi-parser, a prepositional phrase attachment specialist, a conjunct identifier for coordinate conjunctions, and a restructurer. The tagger is an n-gram based program that currently generates syntactic/semantic tags for the words in the corpus. The syntactic portion of the tag is mandatory and the semantic portion depends upon whether the word has any special domain-specific classification or not. Currently only nouns, gerunds, and adjectives are assigned semantic tags. For example, in the domain of veterinary medicine, "dog" would be assigned the tag "noun—patient," "nasal" would be "adj—body-part," etc.

The parsing strategy is based on the initial identification of simple phrases, which are later converted to deeper structures with the help of separate specialist programs for coordinate conjunct identification and prepositional phrase attachment. The result for a given sentence is a single parse, many of whose elements are comparatively underspecified. For example, the parses generated lack clause boundaries. Nevertheless, the results are surprisingly useful in the extraction of relationships from the corpus.

The semi-parser recognises noun-, verb-, prepositional-, gerund-, infinitive-, and adjectival-phrases. The prepositional phrase attachment specialist [2] uses case grammar analysis to disambiguate the attachments of prepositional phrases and is highly domain-dependent. The current implementation of this subcomponent is highly specific to the domain of veterinary medicine, the initial testbed for the KUDZU system. Note that all the examples presented in this paper will be taken from this domain. The coordinate conjunction specialist identifies pairs of appropriate conjuncts for the coordinate conjunctions in the text and is domain-independent in nature [1]. The restructurer puts together the information acquired by the specialist programs in order to provide a better (and deeper) structure to the parse.

Before being passed to the knowledge analysis portion of the system, some parses undergo manual modification, which is

[*] This research is supported by the NSF-ARPA grant number IRI 9314963.

facilitated by the help of an editing tool especially written for this purpose. A large percentage of the modifications can be attributed to the limitation of the conjunct identifier in recognizing only pairs of conjuncts, as opposed to all conjuncts of coordinate conjunctions.

The KA component receives appropriate parses from the NLP component and uses them to populate an object-oriented knowledge base [4]. The nature of the knowledge base created is dependent on a domain schema which specifies the concept hierarchy and the relationships of interest in the domain. Examples of the concept hierarchy specifications and relationships present in the domain schema are given in Figure 1.

```
(class (name patient)
      (parent animal))


(relationship (name treatment)
    (role (name disease) (type mandatory) (class disorder))
    (role (name treatment) (type mandatory) (class PROCEDURE)
                                                  (class MEDICATION))
    (role (name species) (type optional) (class PATIENT))
    (role (name location) (type optional) (class BODY-PART)))
```

Figure 1: Examples of Schema Entries

Many such relationships may be defined in the domain schema. While processing a sentence, the KA component hypothesizes the types of relationships that may be present in it. However, before the actual relationship is instantiated, objects corresponding to the mandatory slots must be found, either directly or by the help of an algorithm that resolves indirect and implied references. If objects corresponding to the optional slots are found, then they are also filled in.

Currently, the domain schema has to be generated manually after a careful evaluation of the domain. This is a time-consuming process that often requires the help of a domain expert. Once the schema has been specified, the rest of the KA component is domain independent [4], with the exception of a domain-specific synonym table. For each sentence that is processed by the KA component, a set of semantic relationships that were found in the sentence is produced. An interface to the KA component allows users to navigate through all instances of the different relationships that have been acquired from the corpus. Two sample sentences from the veterinary medicine domain, their parsed output and the relations extracted from them are shown in Figure 2.

3. OUTLINE OF THE PROPOSED APPROACH

The automatic acquisition of semantic features of a domain is an important issue, since it assists portability by reducing the amount of human intervention. In the context of the KUDZU system in particular, it is desired that the system be moved from the domain of veterinary medicine to that of physical chemistry. As explained above, certain components of the system are domain-dependent and have to be significantly modified before the system can be used for a new domain. The current research aims to use the acquired semantic features in order to improve the portability of the KUDZU system.

It is important to note that although the initial motivation for this research came from the need to move the KUDZU system to a new domain, the underlying techniques are generic and can be used in a variety of applications. The primary goal is to acquire the semantic features of the domain with minimal human intervention, and ensure that these features can be applied to different systems rather easily. In this research, two main types of semantic features are of interest — a concept hierarchy for the domain, and lexico-semantic patterns present in the domain. These patterns are similar to what are also known as "selectional constraints" or "selectional patterns" [6, 5] in systems which use them primarily to determine the correct parse from a large number of parses generated by a broad coverage grammar. They are basically co-occurrence patterns between *meaningful*[1] nouns, gerunds, adjectives, and verbs. For example, "DISORDER of BODY-PART", "MEDICATION can be used to TREAT-VERB PATIENT", etc. are legitimate lexico-semantic patterns from the veterinary medicine domain.

The steps involved in the acquisition of semantic features from the domain can be briefly outlined as follows:

1. Generate the syntactic tags for all the words in the corpus.
2. Algorithmically identify the explicit semantic clusters that may exist in the current domain. Apply the clustering algorithm separately to nouns, gerunds, adjectives, and verbs.
3. Use the syntactic tags and semantic classes to automate the identification of lexico-semantic patterns that exist in the given domain.

This basic methodology is similar to some other approaches adopted in the past [8, 6]. However, some important differences exist which will be discussed later.

Once the semantic features have been obtained, they can be used in a variety of ways depending upon the needs of the NLP system. They can be helpful in improving the portability of an NLP system by providing useful semantic information that may be needed by different components of the system. In the KUDZU system, these features will be used to improve the success rate of a domain-independent syntactically based prepositional phrase attachment specialist, and for automatic acquisition of the domain schema.

It is easy to see how the lexico-semantic patterns can be help-

[1] As explained later, a meaningful word is one that has a semantic tag associated with it.

Sentences:

Cataracts may accompany corneal opacification.
In the Labrador Retriever, they may be associated with skeletal dysplasia of the forelegs.

Parses:

```
(sentence
   (noun_phrase ((w cataracts noun|plural||disorder)))
   (verb_phrase ((w may aux) (w accompany verb)))
   (noun_phrase ((w corneal adj||body_part) (w opacification noun||disorder))))

(sentence
    (prep_phrase (w in prep)
           (noun_pharse ((w the det) (w Labrador noun) (w Retriever noun))))
    (noun_phrase ((w they pro|plural)))
    (verb_phrase ((w may aux) (w be aux) (w associated verb))
           (prep_phrase (w with prep)
                   (noun_phrase ((w skeletal adj||body-part) (w dysplasia noun||disorder))
                           (prep_phrase (w of prep)
                                   (noun_phrase ((w the det) (w forelegs noun|plural||body-part)))))))
```

Relationships:

Relationship:	Symptom	Relationship:	Predisposition
Role SYMPTOM:	cataracts	Role DISEASE:	dysplasia (skeletal)
Role DISORDER:	opacification (corneal)	Role PREDISPOSED:	cataracts
		Role SPECIES:	Labrador Retriever
		Role LOCATION:	forelegs

Figure 2: Sample Sentences Processed by KUDZU

ful in the attachment of prepositional phrases. All patterns that have some preposition embedded within them will essentially provide selectional constraints for the classes of words that may appear in the object and host slots. These patterns will be used to improve the success rate of a domain-independent syntactically based prepositional phrase attachment specialist. There is ample evidence [10, 15] that semantic categories and collocational patterns can be used effectively to assist in the process of prepositional phrase attachment.

These semantic features will also be used to automatically generate the domain schema for any given domain. Figure 1 contains examples of the semantic class hierarchy and the relationships of interest, as defined in the domain schema. The former may be acquired from the semantic clustering process. The specification of the relationships can be achieved with the help of the weighted lexico-semantic patterns. Some of the relationships can be acquired by an automated comparison of all patterns involving a given semantic verb class. Other relationships may be determined by comparing other patterns with common noun and gerund semantic classes. The resulting domain schema, in some sense, represents the semantic structure of the domain.

4. ACQUISITION OF SEMANTIC FEATURES

4.1. Tagging

It was decided that the tag set used by the Penn treebank project, with a few exceptions, be adopted for tagging the corpus. Unlike the Penn treebank tag set, we have separate tags for auxiliaries, gerunds, and subordinate conjunctions (rather than clumping subordinate conjunctions with prepositions). Therefore, as a first step in the process of acquisition of semantic features, the corpus is tagged with appropriate tags. Brill's rule-based tagger [3] is being used for this purpose. This step is primarily domain-independent, although the tagger may have to be further trained on a new domain. Since this tag set is purely syntactic in nature, the semantic clusters of the words must be acquired by a different method.

4.2. Identification of Semantic Clusters

The identification of such semantic clusters (which provide the concept hierarchy) is the next step. Nouns, gerunds, adjectives, and verbs are to be clustered into separate semantic hierarchies. A traditional clustering system — COBWEB/3 — is used to cluster words into their semantic categories. Since COBWEB/3 [13] requires attribute-value vectors associated with the entities to be clustered, such vectors must be defined. The attributes used to define these vectors should be chosen to reflect the lexico-syntactic context of the words because the semantic category of a word is strongly influenced by the context in which it appears. The proposed methodology involves specifying a set of lexico-syntactic attributes separately for nouns, gerunds, adjectives, and verbs. Presumably, the syntactic constraints that affect the semantic category of a noun are different from those that affect the category of gerunds, adjectives and verbs. Currently, three attributes are being used for noun clustering — $subj_{verb}$ (verb whose subject is the current noun), obj_{verb} (verb whose object is the current noun), and $host_{prep}$ (preposition of which the current noun is an object). The top i values that satisfy the attribute $subj_{verb}$, top j values of obj_{verb}, and top k values of $host_{prep}$ are of interest. A cross-product of these values yields the attribute-value vectors. For example, if $i = 3$, $j = 3$, and $k = 2$ are used, $3 \times 3 \times 2 = 18$ vectors are generated for each noun. These values are generated by a program from the phrasal structures produced by the semi-parser. The same attributes can be used across different domains, and hence the attribute-value vectors needed for semantic clustering can be generated with no human intervention. Some examples of the semantic clusters that may be identified in the domain of veterinary medicine are DISORDER, PATIENT, BODY-PART, MEDICATION, etc. for nouns; DIAGNOSTIC-PROC, TREATMENT-PROC, etc. for gerunds; DISORDER, BODY-PART, etc. for adjectives; CAUSE-VERB, TREAT-VERB, etc. for verbs.

The clustering technique is not expected to generate completely correct clusters in one pass. However, the improperly classified words will not be manually reclassified at this stage. In order to attain proper hierarchical clusters, the process of clustering may have to be performed again *after* lexico-semantic patterns have been discovered by the process described below. The only human intervention required at the present stage is for the assignment of class identifiers to the generated classes. In fact, a human is shown small sub-clusters (each with 8-10 objects[2]) of the generated hierarchy, and is asked to label these sub-clusters with a semantic label, if possible. Note that not all such sub-clusters should have semantic labels — several nouns in the corpus are generic nouns that cannot be classified into any semantic class. However, a majority of the sub-clusters should represent the semantic classes that exist in the domain. The class identifiers thus assigned are then associated as semantic tags with these words and used to discover the lexico-semantic patterns in the next step. Any word that has a semantic tag is considered to be *meaningful*.

4.3. Discovery of Lexico-Semantic Patterns

The semantic clusters obtained from the clustering procedure, after the manual assignment of class identifiers, are used to identify the lexico-semantic patterns. The phrase-level parsed structures produced by the semi-parser are analysed for different patterns. These patterns are of the form subject-verb-object, noun-noun, adjective-noun, NP-PP, and VP-NP-PP, where NP, VP, and PP refer to noun-, verb-, and prepositional-phrases respectively. All patterns that occur in the corpus more often than some pre-defined threshold are assumed to be important and are saved. One restriction currently being placed on these patterns is that at least two meaningful words must be present in every pattern. The patterns are weighted on the basis of their frequency of occurrence, with the more frequent patterns getting higher weights.

It seems reasonable to assume that if lexico-semantic patterns were already known to the system, the identification of semantic categories would become easier and vice-versa. In this research, we propose to first identify semantic categories and then the patterns. It has long been realised that there is an interdependence between the structure of a text and the semantic classification that is present in it. Halliday [7] stated that " ... there is no question of discovering one before the other." We believe that an approximate classification can be achieved before the structures are fully identified. However, this interdependence between classification and structure will have its adverse effects on the results. It is anticipated that the results of both semantic clustering and pattern discovery will not be very accurate in the first pass. Therefore, an iterative scheme is being proposed.

After semantic clustering has been performed, human intervention is needed to assign class identifiers to the generated clusters. These identifiers assist in the proper discovery of lexico-semantic patterns. The resulting set of patterns may contain some irrelevant patterns and human intervention is needed to accept/reject the automatically generated patterns. Both the accepted and rejected patterns are stored by the system so that in future iterations, the same patterns do not need human verification. As has been shown before [8, 6, 14], such patterns place constraints on the semantic classes of words appearing in particular contexts. The set of selected patterns can, therefore, be used to reanalyse the corpus in order to recognise the incorrectly clustered words in the previously generated class hierarchy and to suggest the correct class for these words. For example, if the word "penicillin" is incorrectly clustered as a DISORDER, an analysis of the corpus will show that it appears most frequently as a MEDICATION in patterns like "TREAT-VERB DISORDER with MEDICATION", "MEDICATION can be used to TREAT-VERB PATIENT", etc. and rarely as a DISORDER in the DISORDER patterns. Hence, its semantic category can be guessed to be MEDICATION. This guess is added to the list of attributes for the words, and semantic clustering is performed again. This iterative mechanism assists clustering in two ways — firstly, the additional attribute helps convergence towards better clusters and secondly, the tentative semantic classes from the i^{th} iteration can be used to generate values for attributes for the $(i+1)^{th}$ iteration[3], thus reducing the sparsity of data. This time better clusters should be formed, and these will again be used to recognise lexico-semantic patterns. We expect the system to converge to a stable set of clusters and patterns after a small number of iterations. A simple diagram outlining the process of

[2] Note that this does not reflect the size of the final clusters generated by the program, since words that eventually should belong to the same cluster may initially be in different sub-clusters.

[3] When attempting to cluster nouns, for example, the semantic classes for gerunds, verbs, and adjectives are used.

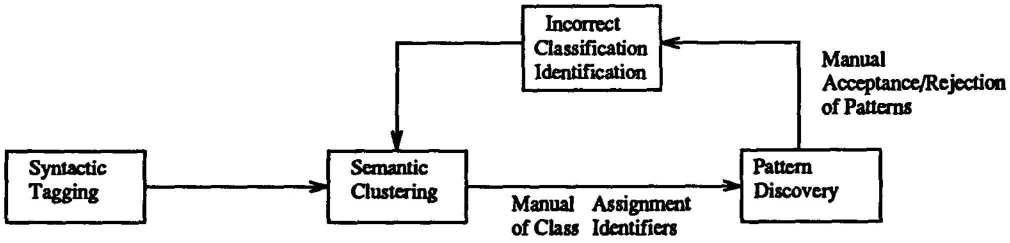
Incorrect Classification Identification
Manual Acceptance/Rejection of Patterns
Syntactic Tagging
Semantic Clustering
Pattern Discovery
Manual Assignment of Class Identifiers

Figure 3: The Proposed Approach

acquisition of semantic features is given in Figure 3.

5. COMPARISON TO OTHER APPROACHES

The most significant work that is similar to the proposed methodology is that conducted on the Linguistic String Project [8] and the PROTEUS project [6] at New York University. Recent efforts have been made by Grishman and Sterling [6] towards automatic acquisition of selectional constraints. The technique proposed here for the acquisition of semantic features should require only limited human intervention. Since the semi-parsed structures can directly be used to generate the attribute-value vectors needed for clustering, the only human intervention should be in assignment of class identifiers, acceptance/rejection of the discovered patterns, and reclassification of some concepts that may not get correctly classified even after the feedback from the lexico-semantic patterns. Further, their approach [8, 6] uses a large broad coverage grammar, and often several parses are produced for each sentence. The basic parsing strategy adopted in the KUDZU system starts with simple phrasal parses which can be used to acquire the semantic features of the domain. These semantic features can then be used for disambiguation of syntactic attachments and thus in providing a better and deeper structure to the parse.

Sekine at al. [16] have also worked on this idea of "gradual approximation" using an iterative mechanism. They describe a scenario for the determination of internal structures of Japanese compound nouns. They use syntactic tuples to generate collocation statistics for words which are then used for clustering. The clusters produced by their program are much smaller in size (approximately 3 words per cluster) than the ones attempted in our research. We intend to generate much larger clusters of words that intuitively belong to the same semantic category in the given domain. The semantic categories generated by the clustering process are used for the identification of semantic relationships of interest in the domain. Most of the emphasis in the research undertaken at New York University on selectional constraints [8, 6] and that in Sekine's work has been on using the collocations for improved syntactic analysis. In addition to using them to disambiguate prepositional phrase attachments, we will also use them to generate a domain schema which is fundamental to the knowledge extraction process.

Our approach of consolidating several lexico-semantic patterns into frame-like structures that represent the semantic structure of the domain is similar to one discussed by Marsh [12]. Several other efforts have been made towards using semantic features for domain-specific dictionary creation or parsing.

6. FINAL COMMENTS

The methodology described in this paper should be useful in acquiring semantic features of a domain with limited human intervention. We also believe that our parsing methodology and the mechanisms for semantic feature acquisition lend themselves very nicely to the development of a simpler and smaller NLP system. This is in contrast to NLP systems that are very large and often use large broad coverage grammars that may take several thousand person-hours to build. The simplicity behind starting with phrasal parses and then using these parses to acquire semantic information that leads to better and deeper parses makes our approach a good "poor man's alternative". The KUDZU system has demonstrated that this simple approach can also yield reasonably good results, at least for data or information extraction tasks.

Acknowledgements

The author is thankful to Dr. Lois Boggess for carefully editing an earlier version of this paper. Thanks are also due to Eric Brill for his tagger and to Kevin Thompson and NASA Ames Research Center for making COBWEB/3 available for this research. Finally, thanks are due to ARPA and NSF for providing the funding for this research.

References

1. Rajeev Agarwal and Lois Boggess. A simple but useful approach to conjunct identification. In *Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics*, pages 15–21. Association for Computational Linguistics, 1992.
2. Lois Boggess, Rajeev Agarwal, and Ron Davis. Disambiguation of prepositional phrases in automatically labelled technical text. In *Proceedings of the Ninth National Conference on Artificial Intelligence*, pages 155–159. The AAAI Press/The MIT Press, 1991.
3. Eric Brill. A simple rule-based part of speech tagger. In *Proceedings of the Speech and Natural Language Workshop*, pages 112–116, February 1992.
4. Jose Cordova. *A Domain-Independent Approach to the Extraction and Assimilation of Knowledge from Natural*

Language Text. PhD thesis, Mississippi State University, August 1992.

5. Ralph Grishman, Lynette Hirschman, and Ngo Thanh Nhan. Discovery procedures for sublanguage selectional patterns: Initial experiments. *Computational Linguistics*, 12(3):205–215, July-September 1986.

6. Ralph Grishman and John Sterling. Smoothing of automatically generated selectional constraints. In *Proceedings of the ARPA Workshop on Human Language Technology.* Morgan Kaufmann Publishers, March 1993.

7. M. A. K. Halliday. Categories of the theory of grammar. *Word*, 17(3):241–292, 1961.

8. Lynette Hirschman. Discovering sublanguage structures. In Ralph Grishman and Richard Kittredge, editors, *Analyzing Langauage in Restricted Domains: Sublanguage Description and Processing*, chapter 12, pages 211–234. Lawrence Erlbaum Associates, Hillsdale, New Jersey, 1986.

9. Lynette Hirschman, Francois-Michel Lang, John Dowding, and Carl Weir. Porting PUNDIT to the resource management domain. In *Proceedings of the Speech and Natural Language Workshop*, pages 277–282, Philadelphia, PA, February 1989.

10. Donald Hindle and Mats Rooth. Structural ambiguity and lexical relations. *Computational Linguistics*, 19(1):103–120, March 1993.

11. Robert Ingria and Lance Ramshaw. Porting to new domains using the learner. In *Proceedings of the Speech and Natural Language Workshop*, pages 241–244, Cape Cod, MA, October 1989.

12. Elaine Marsh. General semantic patterns in different sublanguages. In Ralph Grishman and Richard Kittredge, editors, *Analyzing Language in Restricted Domains: Sublanguage Description and Procesing*, chapter 7, pages 103–127. Lawrence Erlbaum Associates, Hillsdale, New Jersey, 1986.

13. Kathleen McKusick and Kevin Thompson. COBWEB/3: A portable implementation. Technical report, NASA Ames Research Center, June 1990.

14. Phil Resnik. Semantic classes and syntactic ambiguity. In *Proceedings of the ARPA Workshop on Human Language Technology.* Morgan Kaufmann Publishers, March 1993.

15. Phil Resnik and Marti Hearst. Structural ambiguity and conceptual relations. In *Proceedings of the Workshop on Very Large Corpora: Academic and Industrial Perspectives*, pages 58–64, June 1993.

16. Satoshi Sekine, Sofia Ananiadou, Jeremy Carroll, and Jun'ichi Tsujii. Linguistic knowledge generator. In *Proceedings of the 15th International Conference on Computational Linguistics (COLING-92)*, pages 56—566, 1992.