Acquisition of Semantic Patterns for Information Extraction from Corpora *

Jun-Tae Kim and Dan I. Moldovan

Department of Electrical Engineering-Systems
University of Southern California
Los Angeles, CA 90089-2562

Abstract

The memory-based parsing or pattern-based phrase recognition has been successfully applied to the information extraction from natural language corpus. One problem of such an approach is that the parser needs a very large number of semantic patterns. Manual creation of the semantic patterns is unrealistic. To solve the scalability and portability problem, automatic acquisition of semantic patterns must be provided. In this paper, a semi-automatic knowledge acquisition system, PALKA, is presented, which facilitates the construction of a large knowledge base of semantic patterns. The system acquires semantic patterns from texts with a small amount of user interaction. The acquisition process with PALKA is described in detail, and a preliminary experimental result is discussed.

1 Introduction

With the increasing recognition that intelligent behavior depends centrally on large, well-developed, and varied sources of knowledge, several new paradigms have recently emerged in Artificial Intelligence research, including *case-based* and *memory-based* approaches [13] [14] [16]. In these approaches, active memory plays a key role in hosting and guiding the system's reasoning process. The memory-based approach has also been adapted to natural language processing in various ways [7] [12]. In memory-based natural language processing, the system has pre-defined memory structures for various syntactic, semantic and contextual knowledge, and this information is used by the analyzer to interpret the input sentence.

Similar approaches have also been successfully applied to information extraction from texts in the MUC tasks[1]. To extract necessary information efficiently, several successful systems used domain specific knowledge about text patterns, such as phrasal patterns [5], concept sequences [10], or concept nodes [8]. By using such knowledge, or pre-defined memory, one can achieve fast and efficient processing of text by concentrating only on the relevant part of the text and by eliminating the need for full syntactic analysis.

However, one significant problem with the memory-based approach to natural language processing is that it needs a large knowledge base of domain-dependent semantic or phrasal patterns to be created. The creation of such a large body of knowledge is very time consuming, and can be almost impossible for large scale applications. There is also a portability problem when the domain is changed. To solve such problems, an acquisition system which can extract semantic patterns automatically from a large corpus of sentence examples must be provided.

In this paper, a knowledge acquisition tool to extract semantic patterns for memory-based information retrieval system is presented. The major goal of this tool is to facilitate the construction of a large knowledge base of semantic patterns. It acquires new phrasal patterns from the input text, maps each element of the pattern to a meaning frame, generalizes the acquired pattern, and merges it into the current knowledge base. Interaction with user is introduced on some decision points, where the ambiguity can not be resolved automatically without other pieces of predefined knowledge.

Section 2 describes the acquisition system PALKA and the representation of semantic patterns. In section 3, the acquisition procedure is described in detail with examples. The experiment result and future work are discussed in section 4.

*This research has been funded by the National Science Foundation Grant No. MIP 9009109.

[1]Message Understanding Conference sponsored by DARPA. The task is to extract information on terrorist incidents from news articles.

2 The Acquisition System PALKA

The information extraction task is slightly different from the text understanding task. There is a predefined target representation to which the information should be mapped. The goal of the parser is to analyze only the relevant part of the input text and produce a meaning representation which can be easily converted to the target representation. To do that, a large body of domain dependent knowledge on semantics is required. In this section, the representation of semantic patterns for the information extraction and the acquisition system PALKA (Parallel Automatic Linguistic Knowledge Acquisition system) is described.

2.1 Representation of semantic pattern

In our memory-based parsing approach, linguistic knowledge is represented as phrasal patterns in the semantic network. The input sentence is matched against stored patterns in the knowledge base by memory search. The stored pattern consists of the meaning frame and the phrasal pattern pair, and it is called *FP-structure* (Frame-Phrasal pattern structure). The knowledge base is organized as a network of FP-structures and a concept hierarchy.

Figure 1 shows an example of an FP structure. A semantic frame is represented by a set of slots and their semantic constraints on fillers. A phrasal pattern is given by the ordered combination of concepts and lexical entries. An FP-structure is the combination of the two. To combine the phrasal pattern and the frame, each slot of the frame is linked to the corresponding element in the phrasal pattern. The input words are connected to each element in the FP-structure through the *isa* hierarchy of concepts. To interpret a sentence, the parser tries to match the sentence to one of the FP-structures in the knowledge base. When the parsing succeeds, an instance of the FP-structure is generated as a result of the recognition process. More details of the parsing procedure can be found in [10].

As one can see in the example, the meaning of the phrase - or the category of the event - cannot be simply recognized by the main verb. There can be many different domain dependent expressions for the **BOMBING** event, and such patterns can only be acquired by looking at the actual domain corpus.

2.2 The knowledge source

Two major knowledge sources used for acquisition are the *text* and the *template*. A *text* is a set of natural

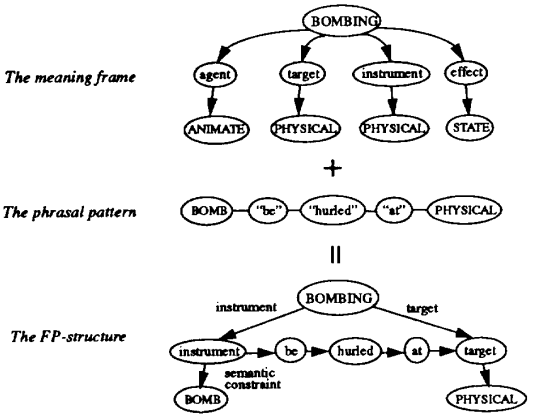

Figure 1: The frame-phrasal pattern representation

language sentences describing domain specific events. The domain currently used is concerned with terrorist events in Latin America, and the text is a set of news articles. PALKA uses the text to acquire phrasal patterns. A *template* is a desired output representation of a sample text which is generated by hand[2]. It contains all the information that should be extracted from the text. If templates are available, PALKA maps the phrasal pattern to the frame definition automatically by using the information in the template. When templates are not available, PALKA acquires the mapping information by comparing the semantic constraint of each slot with the semantic category of each element of the phrasal pattern. User interaction is necessary at some decision points. Other knowledge sources used by PALKA are:

- The *concept hierarchy*, which contains general classification of objects, events and states, and domain specific concepts. It is used to specify a semantic constraint of each element in an FP-structure. After a semantic pattern is acquired, the concept hierarchy is also used for generalization and specialization of the pattern.

- The *frame definition*, which represents the information to be extracted from the domain texts. One of the slots in the definition contains the keyword list, which is used to extract sentences possibly relevant to that frame.

[2]Currently, 1400 news articles on the terrorist domain and their corresponding templates are available on line.

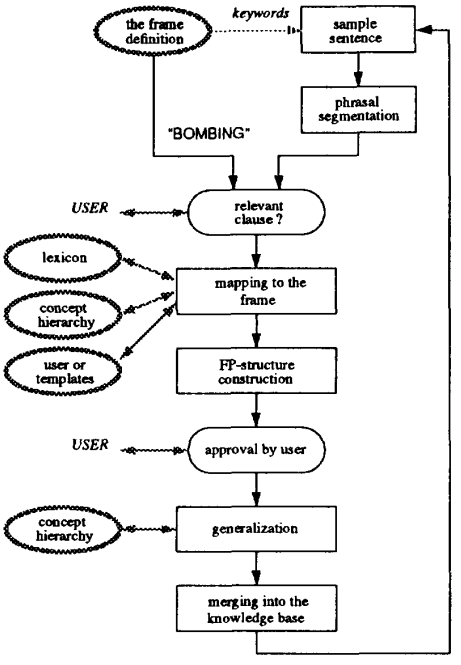

Figure 2: The interactive acquisition procedure

- The *dictionary*, which maps an input word to one or more concepts in the concept hierarchy. For example, "dynamite" is mapped to concept **BOMB**, and "temple" is mapped to concept **BUILDING**, which is linked to concept **PHYSICAL** through *isa* relations.

2.3 Overview of the system

PALKA is a semi-automatic semantic pattern acquisition tool. It acquires domain dependent semantic patterns corresponding to a frame definition which is pre-defined as a target representation for the application domain. The acquired patterns are used for the memory-based information extraction from natural language corpus. It is semi-automatic in the sense that, when the template (target representation) of the given sentence is not available, the user has to interact with the system to determine the mapping of slots to the phrasal pattern. User interaction is also needed to determine the relevancy of a phrase before the system starts the acquisition process, and to determine the correctness of the final form of the FP-structure. In any case, the user's response is very simple, such as a yes/no answer or selecting one of several candidates the system has generated.

Figure 2 shows the functional structure of the acquisition system PALKA. For a given frame definition, the acquisition system selects candidate sentences which may have relevancy, and extracts phrasal patterns from the sentence by segmentation. A simple phrasal parser is used for the segmentation of sentence. It groups the words and converts a complex sentence into several simple clauses. After the segmentation, the system maps each slot of the frame to an element of the phrasal pattern, based on the semantic constraints of the slot and the element's semantic category (concept mapped by the lexicon). When multiple choices exist, selection is made by the user, or by the system using the corresponding template if it exists. The FP-structure is constructed by eliminating unmapped noun groups and prepositional phrases, and by replacing mapped noun groups with their semantic category.

Since the semantic category of the mapped noun group is determined to be the most specific one, it should be generalized if possible. PALKA tries generalization by searching for a compatible FP-structure that has been previously generated. If it exists, the semantic constraints of the same slots are generalized as their common subsumer in the concept hierarchy. The generalized FP-structure is merged into the current knowledge base. While merging, it is combined with other patterns to save space and save parsing time.

3 Acquisition of Semantic Patterns with PALKA

The acquisition of semantic patterns is performed for one frame at a time. For example, the system first acquires all the patterns for **BOMBING** event frame, and then for **KILLING** frame, and so on. In this section, the acquisition procedure in PALKA is described step by step by using the **BOMBING** frame example. The **BOMBING** frame is defined as follows:

```
(BOMBING
   :isa        (TERRORIST-ACTION)
   :keyword    (bomb bombing explode explosion explosive)
   :agent      (ANIMATE)
   :target     (PHYSICAL)
   :instrument (PHYSICAL)
   :effect     (STATE))
```

The first slot, *isa*, points to a more general frame in the knowledge base to which this frame is connected. In the second slot *keyword*, several keywords are specified. Relevant sentences are extracted from the sample texts by using these keywords. The other 4 slots

- *agent*, *target*, *instrument*, and *effect* - indicate the types of information used in this domain. For each slot, a semantic constraint is specified.

Step 1: Keyword based sentence extraction
By using the keyword given in the frame definition, possibly relevant sentences are extracted from the domain corpus. The acquisition procedure is performed for each of the collected sentences. One example of the extracted sentence is shown below. The keyword **bomb** is used to extract this sentence.

```
POLICE HAVE REPORTED THAT AT 0415 THIS MORNING
INCENDIARY BOMBS WERE HURLED AT A MORMON TEMPLE
IN SANTIAGO.
```

Step 2: Phrasal Segmentation
The original text consists of complex sentences which contain relative clauses, nominal clauses, conjunctive clauses, etc. Since semantic patterns are acquired from simple clauses, it is necessary to convert a complex sentence to a set of simple clauses. A simple phrasal parser converts the extracted sentence into simple clauses by words-grouping, simplification, and decomposition of the input sentence. The result of the phrasal parser is:

```
1: [POLICE] [REPORT] [IT]
2: [AT] [0415] [THIS MORNING] [INCENDIARY BOMB]
   [BE HURLED] [AT] [MORMON TEMPLE] [IN] [SANTIAGO]
```

Only the second clause is selected and presented to the user because the first clause does not contain any keyword for bombing. When the segmentized phrase is presented to the user, the user decides its relevancy with respect to the current meaning frame. It is just a simple yes-no answer. It is possible that even if a sentence contains a keyword of a certain frame, it may not have the meaning of that frame. The example clause certainly represents a bombing event, so the user should answer "yes". The phrasal parser also maps each group to one or more concepts in the concept hierarchy by looking at the head noun's semantic category. The syntactic and semantic category of each group is as follows:

```
[AT] preposition
[0415] noun-group, TIME
[THIS MORNING] noun-group, TIME
[INCENDIARY BOMB] noun-group, BOMB
[BE HURLED] verb-group
[AT] preposition
[MORMON TEMPLE] noun-group, BUILDING
[IN] preposition
[SANTIAGO] noun-group, GEO-REGION
```

Step 3: Mapping phrasal pattern to frame
With the given meaning frame and the segmentized clause pattern, the acquisition system tries to find out the mapping between the two to construct an FP-structure. The system selects possible candidates for each slot by using the general semantic constraints given to each slot of the frame. For example, the **target** slot has a semantic constraint of **PHYSICAL**, so possible candidates are the "incendiary bomb" and the "mormon temple", since their semantic categories are under the concept **PHYSICAL** in the concept hierarchy. The selected candidates for each slot of the **BOMBING** frame are as follows:

```
agent:      None
target:     1 INCENDIARY BOMB
            2 MORMON TEMPLE
effect:     None
instrument: 1 INCENDIARY BOMB
```

If a template for this sentence is available, PALKA selects one of the candidates for the **target** slot by looking at the template. In case there is no such template available, the user selects one when there are multiple candidates. Because the system already filtered out several elements in the sentence, the user only needs to select the **target** from the "incendiary bomb" and the "mormon temple", and decide whether the "incendiary bomb" is an **instrument** of this bombing event or not. The user selects the "mormon temple" for the **target** slot and the "incendiary bomb" for the **instrument** slot. The following mapping is thus obtained:

```
[AT] [0415] [THIS MORNING]
[(instrument: INCENDIARY BOMB)]
[BE HURL] [AT] [(target: MORMON TEMPLE)]
[IN] [SANTIAGO]
```

Step 4: FP-structure construction
After all the slot fillers are decided, the system constructs an FP-structure based on the mapping information. The basic strategy for constructing an FP-structure is to include the mapped elements and the main verb, and discard the unmapped elements. The system first filters out all the unmapped prepositional phrases such as "at 0415 this morning" and "in Santiago". Then all the noun groups are replaced by their head noun's semantic categories, and all the verbs are replaced by their root forms. After applying several rules of construction, the phrasal pattern acquired from the example sentence becomes:

```
[(instrument: BOMB) BE HURL AT (target: BUILDING)]
```

Since there was no element with semantic category **ANIMATE** (for **agent** slot) or **STATE** (for **effect** slot), the resultant FP-structure has only **target** and

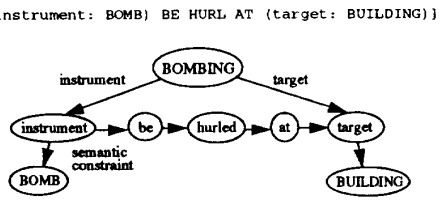

Figure 3: The acquired FP-structure

instrument slots. Figure 3 shows the result.

Step 5: Generalization
The acquired FP-structure is compared with existing ones for further generalization - generalization of each element's semantic constraint. Whenever two FP-structures with the same phrasal pattern are generated, their semantic constraints are generalized by comparing the semantic constraints of same slots and replacing them by their common subsumer in the concept hierarchy.

In our example, the semantic constraint of the **target** slot of the FP-structure is changed from **BUILDING** to **PHYSICAL**. This is because a generalizable pattern with semantic constraint **VEHICLE** is found in another sample sentence, and the most specific common subsumer of **VEHICLE** and **BUILDING** in the concept hierarchy is **PHYSICAL**. The FP-structure after generalization is:

```
[(instrument: BOMB) BE HURL AT (target: PHYSICAL)]
```

The problem of overgeneralization can only be solved by looking at the actual interpretation result. If the parser misinterprets a sentence because of the overgeneralized semantic constraint, the corresponding slot should be specialized (put more specific constraint). Insertion of a trial parsing step within the acquisition procedure for this purpose is discussed in the next section.

Step 6: Merging into the knowledge base
After the acquired FP-structure is appropriately generalized, it is inserted into the knowledge base as a new FP-structure. While inserting, if there is a similar pattern in the knowledge base, the new one is merged with it to save space, and to reduce recognition time when parsing. For example, **[(instrument:BOMB) BE HURL AT (target:PHYSICAL)]** and **[(instrument:BOMB) BE THROW AT (target:PHYSICAL)]** can be merged into one structure.

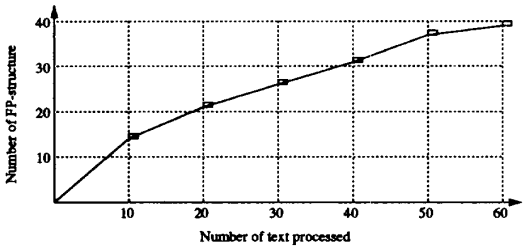

Figure 4: Number of FP-structure vs. text processed

4 Experimental Result

The PALKA prototype has been implemented using C on a SUN workstation, and a preliminary experiment has been performed with 300 MUC texts and the **BOMBING** frame. The domain of the texts is the news articles on terrorist incidents in Latin America. Each text contains an average of 14 sentences. The time spent for the acquisition is less than an hour. Although only one frame and 300 texts were used in the experiment, it is significantly short time compared to several months we spent for the creation of the previous knowledge base for MUC-4 task.

Among the 300 texts, 61 texts contain sentences relevant to the bombing event. From these texts, PALKA acquired a total of 118 phrasal patterns. Among them, 46 patterns were irrelevant or incorrect, 32 patterns were generalized or merged with previous patterns, and 40 new FP-structures were created. The examples of irrelevant patterns are:

```
[FOREIGN DEBT CRISIS] [EXPLODE] [IN] [ANDEAN COUNTRIES]
[EXPLOSION] [OF] [FREEDOM] [COINCIDE] [WITH] ...
[50 KG] [OF] [EXPLOSIVES] [BE FOUND] [INSIDE] [CAR]
```

One of the basic assumptions of our approach to semantic pattern acquisition is that only a finite number of expressions are frequently used in a specific domain, to represent a specific event. In other words, the patterns acquired from a relatively small number of sample texts can cover a much larger number of texts from the same domain. While acquiring knowledge from more and more texts, the acquisition rate should slow down and the number of FP-structures should be saturated. Figure 4 shows this effect. More experiments on 1400 available MUC texts with other frames such as **KILLING**, **ARSON**, **KIDNAPPING**, **ROBBERY**, and **ATTACK** are under preparation. With larger number of training texts, the saturation of the knowledge base will be clearly shown.

As mentioned in the previous section, the memory-based parser should be integrated with the acquisition system to eliminate duplicate acquisition of the same pattern, and to deal with overgeneralization. If there are overgeneralized FP-structures, the parser may recognize the sentence but produce incorrect output representation. For example, if the system generalizes the **target** slot of **[(instrument:BOMB) BE HURL AT (target:X)]** to be the **THING** (which is the most general concept), then the sentence **''THE BOMB WAS HURLED AT 0415 THIS MORNING''** will be recognized as:

```
(BOMBING
   :instrument BOMB
   :target     0415)
```

By introducing a trial parsing step, the overgeneralization can be detected by the user, and it can be corrected by specializing the incorrect slot constraint.

5 Conclusion

The memory-based parsing or pattern-based phrase recognition has been successfully applied to the information extraction from a natural language corpus. However, such a system needs a large knowledge base of semantic patterns which is difficult to create. There are the scalability and the portability problems for practical, large scale applications. In this paper, a semi-automatic semantic knowledge acquisition system PALKA has been presented, by which one can overcome such problems. The goal of the acquisition system is to facilitate the construction of a large knowledge base of semantic patterns. PALKA acquires new phrasal patterns for domain specific frames from sample natural language corpus. The acquired semantic patterns are further tuned through generalizations of their semantic constraints.

An experiment using PALKA with 300 MUC-4 domain texts demonstrates the feasibility of our approach. By using this system, the time to construct the knowledge base of semantic patterns is reduced significantly. The result shows that PALKA can be used successfully to acquire semantic patterns for practical information extraction system.

References

[1] J. D. Becker, "The phrasal lexicon", *Bolt Beranek and Newman Inc. Report No. 3081*, 1975.

[2] R. C. Berwick, *The Acquisition of Syntactic Knowledge*, The MIT press, 1985.

[3] R. J. Brachman and J. G. Schmolze, "An overview of the KL-ONE knowledge representation system", *Cognitive Science*, vol. 9, 1985.

[4] A. Hauptmann, "From syntax to meaning in natural language processing", *Proc. of AAAI-91*, 1991.

[5] J. R. Hobbs, D. Appelt, M. Tyson, J. Bear and D. Israel, "FASTUS: System summary", *Proc. of MUC-4*, 1992.

[6] J.-T. Kim and D. Moldovan, "A knowledge acquisition model for memory-based parsing", *Technical Report PKPL 91-9*, University of Southern California, Department of EE-Systems, 1991.

[7] H. Kitano, "ΦDM-Dialog. An experimental speech-to-speech dialog translation system", *IEEE Computer*, June, 1991.

[8] W. Lehnert, C. Cardie, D. Fisher, J. McCarthy, E. Riloff and S. Soderland, "Description of CIRCUS system as used for MUC-3", *Proc. of MUC-3*, 1991.

[9] D. Moldovan, W. Lee, C. Lin, and M. Chung, "SNAP: Parallel processing applied to AI", *IEEE Computer*, June, 1992.

[10] D. Moldovan, S. Cha, M. Chung, K. Hendrickson, J. Kim and S. Kowalski, "Discription of the SNAP system used for MUC-4 ", *Proc. of MUC-4*, 1992.

[11] M. T. Pazienza and P. Velardi, "Methods for extracting knowledge from corpora", *Proc. of the 5th Annual Workshop on Conceptual Structures*, 1990.

[12] C. K. Riesbeck and C. E. Martin, "Direct memory access parsing", *Report 354*, Dept. of Computer Science, Yale University, 1985.

[13] C. K. Riesbeck and R. Schank, *Inside Case-Based Reasoning*, Lawrence Erlbaum Associates, 1989.

[14] C. Stanfill and D. Waltz, "Toward memory based reasoning", *Communication of ACM*, 29-12, 1986.

[15] P. Velardi, M. T. Pazienza and S. Magrini, "Acquisition of semantic patterns from a natural corpus of texts", *ACM SIGART Newsletter*, No. 108, April 1989.

[16] D. L. Waltz, "Massively parallel AI", *Proc. of AAAI-90*, 1990.

[17] R. Wilenski, Y. Arens and D. Chin, "Talking to UNIX in English: An overview of UC", *Communications of the ACM*, Vol. 27, No. 6, 1984.

[18] U. Zernik, *Strategies in Language Acquisitions: Learning Phrases from Examples in Context*, PhD Dissertation, Computer Science Dept., UCLA, 1987.