A Three-Tiered Approach to Natural Language Text Retrieval *

Tim Finin, Robin McEntire, Carl Weir, and Barry Silk
Unisys Corporation
Center for Advanced Information Technology
Paoli, Pennsylvania

Introduction

This paper presents a three-tiered approach to text processing in which a novel and quite powerful knowledge-based form of information retrieval plays a central role. We are actively using this approach in our efforts for the NOSC-sponsored *Third Message Understanding Conference* (MUC-3) [?]. Research groups participating in the MUC-3 conference must evaluate the performance of their text processing systems on a black-box, template (database record) generation task. To perform this task, a text processing system must extract information about different types of terrorist acts from newspaper articles and radio broadcasts. Relevant data about terrorist acts including when and where they occurred, who perpetrated them, what weapons were used, who or what were the targets, and so forth, constitute the content of templates used to represent them in a database. The knowledge-based form of information retrieval which plays a key role in our three-tiered approach allows us to define an interesting level of text analysis that falls somewhere between what is possible with standard IR techniques and deep linguistic analysis.

Our three-tiered approach to text processing can be defined in terms of three processing components: a keyword analysis system that is used to predict the occurrence of terrorist act descriptions; the knowledge-based information retrieval system KBIRD which is used to instantiate templates for the terrorist act descriptions detected by the keyword analysis system; and a natural language processing system called PUNDIT [?], which KBIRD provides with key segments of text on which to perform a detailed linguistic analysis in order to extract information about grammatical and thematic roles. In the next three sections of this paper, these three components are described in more detail.

*This work was partially supported by DARPA Contract XXX. Submitted to the AAAI-91 Workshop on Natural Language Text Retrieval.

Keyword Analysis

The keyword analysis component of the Unisys MUC-3 system predicts when various types of terrorist acts (bombings, murders, kidnappings, and so forth) have been referred to in a text. The probability of an act of a given type having occurred is determined by a search for words, word stems, and pairs of words and word stems, that are associated with types of acts.[1] The probability of a such a word (or word stem, or word or word stem pair) occurring in a message for which an act of a given type is associated is determined as follows:

1. The frequency of presence for a given word W (or word stem ...) in messages for which a terrorist act of a given type T occurs is computed ($f(W,T)$), as is the presence of the word in any message at all in the complete message corpus ($f(W,C)$).
2. The probability of the word appearing in messages for which a terrorist act of a given type occurs ($\frac{f(W,T)}{f(T,C)}$) and the probability of the word occurring in any message at all ($\frac{f(W,C)}{|C|}$) are calculated, and these two values are used to determine the conditional probability of the word (or word stem ...) predicting the given type of terrorist act.
3. Only words with relatively high probabilities of predicting a given type of terrorist act are searched for in a text, and words that don't occur frequently enough in the text corpus based on some empirically-derived threshold are not used.

Stated more concisely:

$$\mathcal{P}(W,T) = \frac{\left(\frac{f(W,T)}{f(T,C)}\right)}{\left(\frac{f(W,C)}{|C|}\right)},$$

where $\mathcal{P}(W,T)$ is the probability of a word W (or word stem ...) predicting a terrorist act of type T.

[1] The keyword analysis system uses a rule-driven word-stemmer based on one developed by Chris Paice (Landcaster, UK) [?].

A drawback of the keyword analysis component in its current state is that it is unable to predict multiple terrorist acts of the same type within a single text.[2] A way to eliminate this drawback is now being worked on.

KBIRD

Once a set of terrorist acts have been predicted, the task of generating templates describing those acts falls to the knowledge-based information retrieval component we have built called KBIRD.

KBIRD is a rule-based system for concept-spotting in free text. KBIRD rules are forward-chaining horn clauses whose antecedents are constituents discovered and recorded in a chart data structure and whose consequents are newly inferred constituents—concepts (or facts)—to be added to the chart. The antecedents and consequents of KBIRD rules can include arbitrary Prolog goals just as in *Definite Clause Grammars.*

It is tempting to think of a set of KBIRD rules as implementing a kind of bottom up chart parser, but their are several interesting differences. One distinctive feature of KBIRD rules is that the concepts they infer are associated with a specific region of text, a region which is the maximal cumulative span of the regions of text associated with each expression in a given rule's antecedent. Moreover, these regions can be explicitly reasoned about by subsequent KBIRD rules.

In typical natural language parsers, there is an implicit constraint that adjacent constituents in a rule must be realized by contiguous strings of text in the input. KBIRD allows one to write rules which specify other constraints on the relative positions of the strings which realize rule constituents. The antecedent of a KBIRD rule may consist of several facts (words or concepts) that are the arguments of operators of the following sort. New operators are easy to define in KBIRD. The ones provided are just a sampling.

Antecedent Format	**Operator Description**
A ^ B	A is contiguous with B.
A , B	A is in the same message as B.
A .. B	A is in the same sentence as B.
A ... B	A is in the same paragraph as B.
A ..+ B	A is in the same region as B.

KBIRD has many additional features which are inherited from the Pfc [?] rule language which provides the implementation substrate. For example, it is possible to write non-monotonic rules which specify, for example, that no occurence of a certain constituent be found in a given region.

[2] This drawback leads to a high number of false negative predictions.

```
1.  "MURDER*" ⟹ potential_murder_event.

2.  "ARMY" ^ "OF" ^ "NATIONAL" ^ "LIBERATION"
    ⟹ terrorist_organization.

3.  "RICARDO" ^ "ALFREDO" ^ "CASTELLAR"
    ⟹ government_person.

4.  terrorist_event(E) .. potential_victim(V)
    ⟹ victim(E,V).

5.  bombing_event ⟹ terrorist_event(bombing).

6.  peasant ⟹ potential_victim.

7.  government_person ⟹ potential_victim.
```

Figure 1: Examples of KBIRD rules

Some examples of KBIRD rules are shown in Figure ??. The first rule states that if the word-stem "MURDER*" has been found in the text, then a fact should be added to the factbase stating that a potential_murder_event has been found. The second and third rules illustrate KBIRD's ability to recognize phrases. The second rule asserts that if the string "ARMY OF NATIONAL LIBERATION" is discovered, a fact should be added to the factbase stating that a terrorist organization exists in the text at the same location as the string. Similarly, the third rule asserts that if the string "RICARDO ALFREDO CASTELLAR" is discovered, then a fact should be added to the factbase that a government_person has been detected at the same location in the text as the string—a different occurrence of a government_person may occur at some other location in the text. The remaining four rules contain examples of operations on concepts derived from the text. The fourth rule, for example, asserts that if a terrorist event E is found in the same sentence as a potential victim V, then a fact should be added to the factbase indicating that V is the actual victim of E.

PUNDIT

The PUNDIT natural language processing system that has been under development at Unisys for the last five years is capable of performing a detailed linguistic analysis of an input text. Unlike KBIRD, PUNDIT abstracts away from the actual strings used to convey information in a text at the very beginning of its analysis process by determining which syntactic properties and domain concepts the lexical items in the text correspond to. These syntactic properties and domain concepts are then processed without much attention being paid to their physical location in the text. In KBIRD, on the

other hand, everything that is manipulated, even concepts that have been asserted, are explicitly associated with regions of text.

A key capability that the deeper linguistic processing of PUNDIT can provide is the determination of the grammatical and thematic roles of expressions in a text. Thus, it can determine that in the following sentence *Castellar* is the subject of the copular verb in the matrix clause, and that *Castellar* should inherit properties asserted of the predicate nominal argument. It can also recognize the passive voice of the relative/subordinate clause headed by *that* and thus that it is Castellar that has been murdered (as the second mayor) in Columbia.

Castellar is the second mayor that has been murdered in Colombia in the last 3 days.

It would be possible to build a KBIRD rulebase that performs the sort of detailed linguistic analysis now being performed by PUNDIT. Merging KBIRD and PUNDIT in this way would minimize the complications of integrating the text analyses that they perform. However, such a merger would very likely reduce the modularity of the three-tiered approach to text processing that we have been following. We intend to more fully explore KBIRD's capabilities before worrying about striking a proper balance between integrated processing and modularity.

Conclusion

The value of our three-tiered approach is two-fold. First, the domain in which we are currently working is so well-defined that a deep linguistic analysis is rarely needed. Using linguistic analysis sparingly and perhaps not at all in some texts provides a dramatic improvement in processing time. Second, in the MUC-3 evaluation task we have discovered that a small amount of modeling effort, i.e., writing KBIRD rules, produces a significant improvement in our ability to extract pertinent information. Since KBIRD is a forward chaining rule-driven methodology, the creation, modification and removal of rules is a very easy and intuitive process.

The three-tiered approach of combining traditional information retrieval and linguistic analysis techniques with the type of analysis that our knowledge-based information retrieval system, KBIRD, provides offers significant advantages to solving common text processing problems. The modularity of this approach allows us to utilize advances made in keyword analysis and NLP technology with relative ease.

References

[1] T. Finin, R. Fritzson, and D. Matuzsek. Adding forward chaining and truth maintenance to prolog. In *CAIA-89*, pages 123–130, March 1989.

[2] L. Hirschman, M. Palmer, J. Dowding, D. Dahl, M. Linebarger, R. Passonneau, F.-M. Lang, C. Ball, and C. Weir. The PUNDIT natural-language processing system. In *AI Systems in Government Conf.* Computer Society of the IEEE, March 1989.

[3] Chris Paice. Another stemmer. *SIGIR Forum*, Fall 1990.

[4] Beth M. Sundheim. Third message understanding conference (MUC-3): Phase 1 status report. In *Proceedings of the DARPA Speech and Natural Language Workshop*, February 1991.