Duke's Trainable Information and Meaning Extraction System (Duke TIMES) *

Amit Bagga
Joyce Yue Chai
Department of Computer Science
Box 90129, Duke University
Durham, NC 27708-0129
Internet: {amit, chai}@cs.duke.edu

1 Introduction and Background

The explosion in the amount of free text materials on the Internet, and the use of this information by people from all walks of life, has made the issue of generalized information extraction a central one in Natural Language Processing. Many systems including ones from NYU, BBN, SRI, SRA, and MITRE have taken steps to make the process of customizing a system for a particular domain an easy one.

We have built a system that attempts to provide *any user* with the ability to efficiently create and customize, for his or her own application, an information extraction system with competitive precision and recall statistics.

More details about the system can be found in (Bagga, 1997).

2 System Architecture

As illustrated in Figure 1, there are three main stages in the running of the system: the Training Process, Rule Generalization, and the Scanning Process. During the Training Process, the user, with the help of a graphical user interface, takes a *few* prototypical articles from the domain that the system is being trained on, and creates rules (patterns) for the target information contained in the training articles. These rules are specific to the training articles and they are generalized so that they can be run on other articles from the domain. The Rule Generalization routines, with the help of WordNet [1] (Miller, 1990), generalize the specific rules generated by the Training Process. The system can now be run on a large number of articles from the domain (Scanning Process). The output of the Scanning Process, for each article, is a semantic network for that article which can then be used by a Postprocessor to fill templates, answer queries, or generate abstracts.

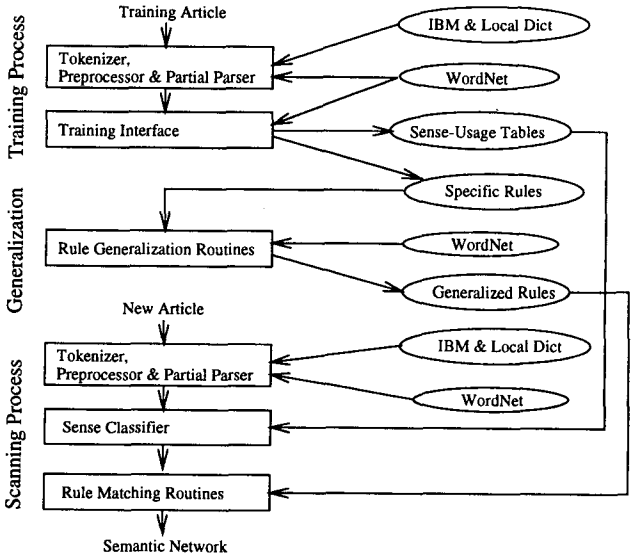

Figure 1: The Architecture

2.1 Tools Used By the System

In addition to WordNet, the system uses IBM's LanguageWare English Dictionary, IBM's Computing Terms Dictionary, and a local dictionary of our choice. The system also uses a gazetteer consisting of approximately 250 names of cities, states, and countries.

2.2 The Tokenizer, the Preprocessor, and the Partial Parser

The Tokenizer accepts ASCII characters as input and produces a stream of tokens (words) as output. It also determines sentence boundaries.

The preprocessor tries to identify some important entities like names of companies, proper names, etc. contained in the article. Groups of words that comprise these entities are collected together and con-

*Supported by Fellowships from IBM Corporation.

[1] WordNet is an on-line lexical reference system developed by George Miller at Princeton University.

sidered as one item for all future processing.

The Partial Parser produces a sequence of non-overlapping phrases as output. The headword of each phrase is also identified. The parser recognizes noun groups, verb groups and preposition groups [2] (Hobbs, 1993).

2.3 The Training Interface

There are two parts to the Training Process: identification of the (WordNet) sense usage of headwords of interest, and the building of specific rules. Training is done by a user with the help of a graphical user Training Interface.

3 Generalization

Rules created as a result of the Training Process are very specific and can only be applied to exactly the same patterns as the ones present during the training. Generalization consists of replacing each concept in a rule by a more generalized concept (obtained from WordNet). Figure 2 shows the different degrees of generalization of the concept "IBM Corporation."

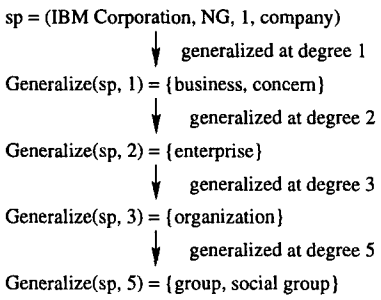

Figure 2: Degrees of Generalization

4 Experiments

We designed an experiment to investigate how training and the generalization strategy affect meaning extraction. We trained our system on three sets of articles from the *triangle.jobs* USENET newsgroup, with emphasis on the following seven facts: Company Name, Position/Title, Experience/Skill, Location, Benefit, Salary, and Contact Information.

The first training set contained 8 articles; the second set contained 16 articles including the first set; and the third set contained 24 articles including those in the first two sets. For rules from each training set, seven levels of generalization were performed. Based on the generalized rules at each level, the system was run on 80 unseen articles from the same newsgroup to test its performance on the extraction of the seven facts.

[2] We wish to thank Jerry Hobbs of SRI for providing us with the finite-state rules for the parser.

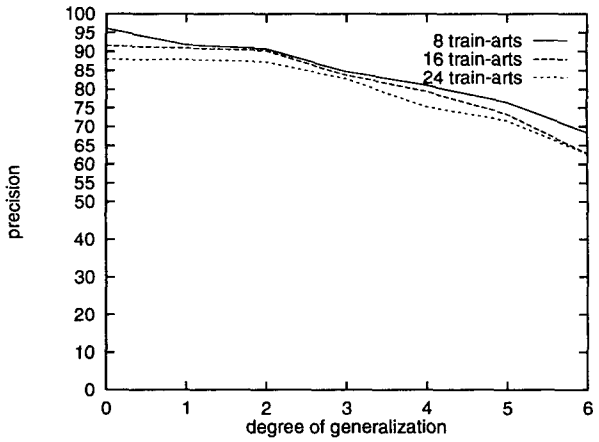

Figure 3: Precision vs. Degree of Generalization

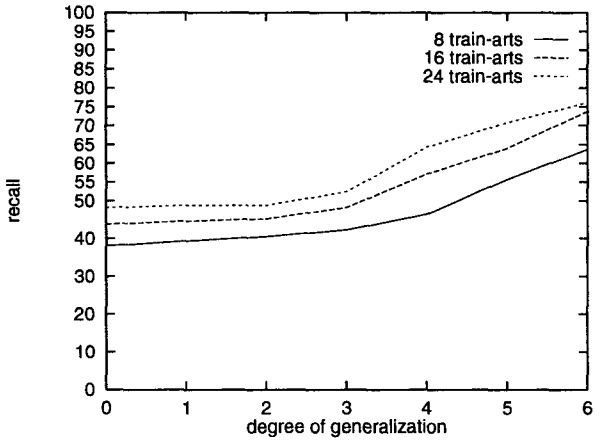

Figure 4: Recall vs. Degree of Generalization

The precision and recall curves with respect to the degree of generalization are shown in Figure 3 and Figure 4 respectively.

References

Bagga, Amit, and Joyce Y. Chai. 1997. A Trainable Message Understanding System, Submitted to the *Fifteenth International Joint Conference on Artificial Intelligence* (IJCAI'97).

Hobbs, J., et al. 1995. FASTUS: A system for Extracting Information from Text, *Human Language Technology*, pp. 133-137, 1993.

Miller, G.A., et al. 1990. *Five Papers on WordNet*, Cognitive Science Laboratory, Princeton University, No. 43, July 1990.