Rule-based Extraction of Experimental Evidence in the Biomedical Domain – the KDD Cup 2002 (Task 1)

Yizhar Regev
ClearForest Ltd
Or Yehuda, Israel
ryizhar@clearforest.com

Michal Finkelstein-Landau
ClearForest Ltd
Or Yehuda, Israel
michal@clearforest.com

Ronen Feldman
ClearForest Corp.
New York, NY, USA
ronen@clearforest.com

ABSTRACT

Below we describe the winning system that we built for the KDD Cup 2002 Task 1 competition. Our system is a Rule-based Information Extraction (IE) system. It combines pattern matching, Natural Language Processing (NLP) tools, semantic constraints based on the domain and the specific task, and a post-processing stage for making the final curation decision based on the various evidence (positive and negative) found within the document. Development and implementation were made using the ***DIAL*** IE language and the ***ClearLab*** development environment. The results achieved were significantly superior than those achieved using categorization approaches.

1. INTRODUCTION

Papers discussing the Drosophila genes and their products are lengthy and complex. They typically include not only text, but also images that are crucial for understanding the papers' results. On the other hand, they usually have a relatively fixed structure and present results obtained using a set of known techniques (such as Northern or Western Blot). Since the focus of this task is the experimental results and not the other information within the paper, focusing on templates used within the figure legends, together with information within the title and the paper abstract proved to be sufficient for most cases. The rest of this paper is organized as follows: in Section 2 we describe briefly our rule-based approach and in Section 3 we outline the actual implementation of the system in the DIAL language. In Section 4 we present the results and discuss briefly the advantages of our approach in comparison to other approaches.

2. RULE-BASED IE APPROACH

IE approaches can be divided into supervised, rule-based approaches and unsupervised (statistical) approaches. The most common unsupervised approach is Hidden Markov Model (HMM), see for example [3]. While unsupervised methods are generally considered less domain-specific and thus theoretically more attractive, actual implementations in complex domains such as the biomedical domain proved to be rather difficult.

We believe that the rule-based approach is more suitable for the KDD Cup task. Presented at the most simplified level, this approach involves writing rules for matching the common patterns for the desired template (experimental evidence for a gene product expression), as in the figure legend: "**Fig 4. Expression of dGATAc transcript**". However, our approach involves more than simple pattern matching. It uses lexical resources, NLP tools and semantic constraints, achieving better coverage and accuracy. In [1] we described this approach in detail in the context of the financial news domain. Below we describe briefly the elements adapted for the KDD Cup task.

2.1 Lexical Resources

The system uses lexicons for key pattern elements such as analysis techniques, positive headline keywords (such as "homologue") and negative headline keywords (e.g. "ectopic", i.e. unnatural).

A general gene lexicon is used, together with a gene thesaurus. In addition, we use a lexicon of the gene candidates listed in the header of each document. One of the functions of this "local" lexicon was to filter noisy entries in the general lexicon. While strings such as "is" won't usually be a gene name, occurrence in the header list makes them more likely to constitute a gene name. We also had to write rules for normalizing typography such as "**Dgc [alpha.gif] 1**" as "**Dgc&agr ; 1**" for the ASCII representation of the Greek letter "alpha".

2.2 NLP Tools

Our system includes three layers of NLP tools, whose function is to extract full and syntactically sound pattern elements:

(i) Part-of-Speech (POS) tagger

(ii) Noun Phrase and Verb Phrase Grouper: Grouping together the head noun with its left modifiers, for example: "**the developing midgut**" and, for verbs, chunking a main verb with its auxiliaries, as in "**does not antagonize**". In the previous examples, "**midgut**" is the head and "**developing**" is a modifier. "**antagonize**" is the main verb and "**does**" is an auxiliary.

(iii) Verb and Noun Pattern Extractor: Extracting larger verb and noun phrases, based on semantic requirements. Example: "**Dac does not antagonize hth expression**". This extractor matches verbs and nouns with their complements. (Here – "**hth expression**" is the complement (the direct object)).

2.3 Semantic Constraints

This was a crucial part of our implementation, since in our task it was critical to know when a gene name is actually part of a transgene, or when a gene expression was not a Wild-Type expression on its own, but rather an evidence for a functional dependency achieved by the researchers ectopically.

Accordingly, if the **hsp70** gene is found within a phrase such as **"@hsp70@-@white@ transgene**", it is ignored. Similarly, if a gene expression phrase is found within a verb phrase that describes a functional dependency result it is also ignored. (as in: "**Dac does not antagonize hth expression in the antenna**").

2.4 Post Processing

The KDD task was unique in that the output required was not single phrases within the document, but a global decision regarding the whole document (whether it should be curated or not). In order to support this decision, the IE process included two main stages:

(i) Extracting evidence from the title, abstract, figure legend and GenBank footnotes, keeping a **score** entry for the whole document and for each product (transcript/protein) of a candidate gene

(ii) Using the scores to decide on the curation of the document and the products of the candidate genes, after processing the document: if a gene's score is above a certain threshold, mark the gene as having an experimental result, and mark the whole document as curatable.

```
DIAL Rule example :

Induced expression - The gene expression is induced or induces another activity
(and is NOT observed on its own), as in : "Fig. 4. Dac does not antagonize hth
expression in the antenna."

//lexicon for relevant nouns similar to "expression":
wordclass wcExpressionNoun = expression transcription localization detection ;

//lexicon for verbs indicating induction/interaction between genes as "antagonize":
wordclass wcInducedVerbs = reduce inhibit activate induce repress alter antagonize ;

//extract Noun Phrase (NG-Noun Group) incorporating a gene
GeneExpressionNG() :-
ExtractedGene(Gene,Product)                    //"Dac" (The gene)
NounGroup(Article,Head,Stem)                   //"expression"
verify(InWC(Head,@wcExpressionNoun)) ;         //verify that the Head is relevant

//Rule for the induced Expression itself
Induced_Expression() :-
ExtractedGene(Gene,Product,mutant)              //"Dac"
VerbGroup(Stem,Tense,Aspect,Voice,Polarity)  //verb group-"does not antagonize"
GeneExpressionNG                                // "hth expression"
verify(InWC(Stem,@wcInducedVerbs));             //verify that the Stem is indeed a
                                                //relevant verb
```

Figure 1 – Sample DIAL Rule

3. IMPLEMENTATION IN DIAL

Our system is implemented in **DIAL** (**D**eclarative **I**nformation **A**nalysis **L**anguage), a rule-based general IE language developed at ClearForest. We outline its main elements below and give an example for a rule in Figure 1. More information is provided in [1].

The "building blocks" of DIAL are rules. Rules are sequences of *Pattern Matching* elements, augmented by a set of *Constraints* that the matched patterns must obey and by a set of *Assignments* of the rule's parameters and/or *actions* concerning external variables / data structures. The Pattern Matching elements themselves can be either: literal strings found in the text (e.g. "*expression*"), a lexicon (called in DIAL wordclass, e.g. *wcInducedVerbs* in Figure 1), or another rule (e.g. *ExtractedGene*). A sample constraint is shown at the end of the *InducedExpression* predicate – *Stem* must be a member (*InWC*) of *wcInducedVerbs.* (Otherwise, it isn't an "Induced Expression").

DIAL enables the user to implement separately the different operations required for performing IE: tokenization, sectioning (recognizing sentence boundaries), and morphological and lexical processing, parsing and domain semantics. DIAL has built-in modules that perform the general tasks of tokenization and part-of-speech tagging. In addition, we have developed a general library of rules that perform Noun Phrase and Verb Phrase grouping. Figure 2 presents ClearForest's development environments for DIAL rules – **ClearLab.** This application enables the user to test the ***Rulebook*** (a set of rule modules) on relevant document collections, view the extracted instances of the templates and their location in the original text. This allows the user to change and correct (debug) the rules as necessary.

4. RESULTS AND EVALUATION

We achieved an F-Measure score of 78% in the Document Curation task and an F-Measure score of 67% in the Gene Product task. These results are significantly better than the results achieved using categorization approaches (we achieved only 62-64% in the Document Curation task using our categorization tool). We believe that the rule-based IE approach was successful for the following reasons:

(i) Most papers use quite a narrow vocabulary.

(ii) Many curatable papers have both relevant results (wild-type expression) and irrelevant ones (mutations etc.)

(iii) Extracting evidence of specific gene products cannot be achieved by categorization. Patterns with the specific genes must be found. There aren't genes that are always relevant and genes that are not, other than ***w***, the "white eye" gene. The training papers never have relevant results for ***w*** products and only mention ***w*** as a tool for studying other genes.

Our IE approach is advantageous also because it provides the maximal information to the user and can be integrated into a full text mining system that allows the user to see the results of the IE process visually and correct / change them as necessary (See [2] for description of such a system).

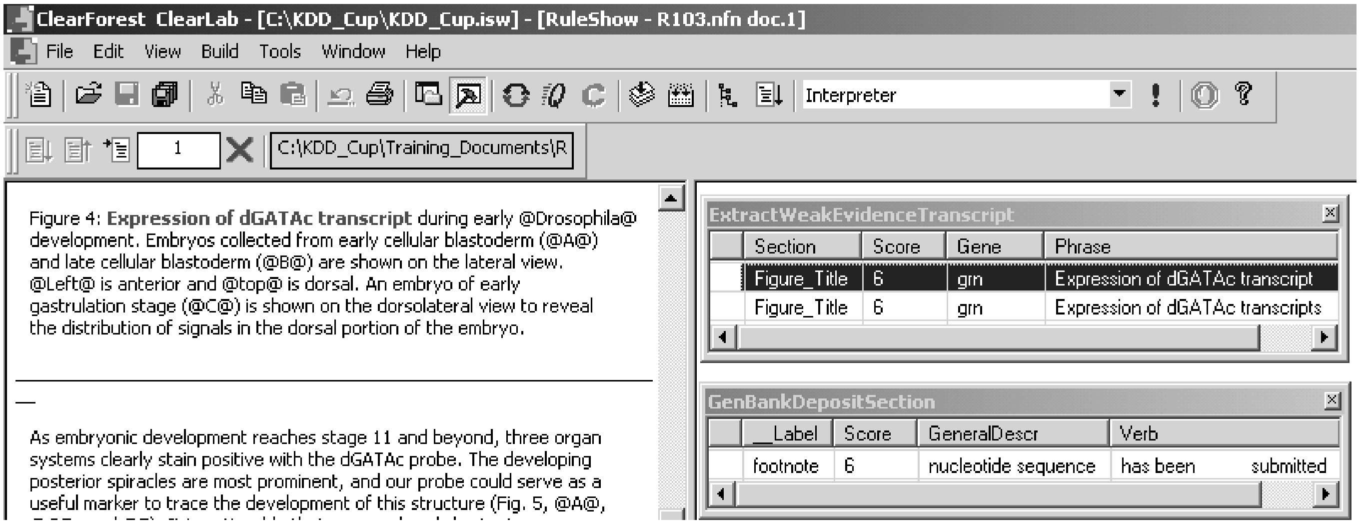

Figure 2 – The ClearLab Application

5. REFERENCES

[1] Feldman, R. et al. A Comparative Study of Information Extraction Strategies. in Proceedings of CICLing 2002 (Mexico City, February 2002), Springer Verlag, 349-359.

[2] Feldman, R. et al. Mining Biomedical Literature using Information Extraction. Current Drug Discovery (October 2002) 19-23.

[3] Ray, S. & Craven, M. Representing Sentence Structure in Hidden Markov Models for Information Extraction. In Proceedings of the 17th International Joint Conference on Artificial Intelligence (IJCAI-2000).