

# A Bayesian network coding scheme for annotating biomedical information presented to genetic counseling clients

Nancy Green\*

*Department of Mathematical Sciences, University of North Carolina at Greensboro, Greensboro, NC 27402-6170, United States*

Received 13 July 2004

Available online 11 November 2004

## Abstract

We developed a Bayesian network coding scheme for annotating biomedical content in layperson-oriented clinical genetics documents. The coding scheme supports the representation of probabilistic and causal relationships among concepts in this domain, at a high enough level of abstraction to capture commonalities among genetic processes and their relationship to health. We are using the coding scheme to annotate a corpus of genetic counseling patient letters as part of the requirements analysis and knowledge acquisition phase of a natural language generation project. This paper describes the coding scheme and presents an evaluation of intercoder reliability for its tag set. In addition to giving examples of use of the coding scheme for analysis of discourse and linguistic features in this genre, we suggest other uses for it in analysis of layperson-oriented text and dialogue in medical communication. © 2004 Elsevier Inc. All rights reserved.

**Keywords:** Content annotation; Content analysis; Intercoder agreement; Bayesian network; Genetic counseling; Clinical genetics; Patient communication; Natural language generation; Natural language processing; Discourse annotation

## 1. Introduction

As information about human genetics has increased rapidly in the last few years, so have genetic testing options such as newborn screening for inherited disorders, testing for genetic predispositions to certain types of cancer, and testing for genetic variations that may determine the effectiveness of certain medications. In the USA, genetic counselors meet with clients to discuss testing options, risk of complications, interpretation of test results, diagnosis of inherited conditions, and recurrence risks. An important function of the counselor is to provide educational counseling, i.e., to provide information in terms that are comprehensible to a lay person. The *patient letter* is a standard document written by a

genetic counselor to her client summarizing the services and information provided to the client [1].

We are analyzing a corpus of patient letters written by genetic counselors to gain knowledge for the design of a computer-supported health communication system for clinical genetics. Employing various artificial intelligence methods, including natural language generation (NLG) techniques [2], the proposed system will automatically produce the first draft of a patient letter using general information on clinical genetics from a knowledge base and documentation about the client's case from her healthcare providers. By generating a first draft, the system could save the counselor time, as well as provide her with easy access to information on related research and patient resources. Note that it is *not* the goal of the proposed system to automate problem-solving tasks such as recommending testing, performing diagnosis, or calculating risk, but rather to provide support to healthcare professionals in performing patient-tailored communication tasks.

\* Fax: +1 336 334 5949.

E-mail address: [nlgreen@uncg.edu](mailto:nlgreen@uncg.edu).

Many NLG-capable systems synthesize texts in a target language, such as English, starting from a nonlinguistic representation of information stored in a knowledge base (KB); also, the systems may separate the task of *discourse planning*, automatically selecting relevant information from a KB and determining how to organize it into structural units of discourse, from the task of *linguistic realization*, automatically synthesizing sentences to convey the selected information [2]. Our analysis of a corpus of clinical genetics patient letters follows a related three-way division. First, we are analyzing the biomedical content of the letters to help design a nonlinguistic KB for the health communication system and to provide an abstraction of the letters' contents to facilitate the next two levels of analysis. Second, we are analyzing (discourse) structural devices in the letters. Third, we are analyzing their (sentential) linguistic features.

Analyzing a corpus of sample documents as one of the first steps towards design of the generation components of a system is a standard methodology in the NLG field for understanding user requirements and characteristics of the target genre<sup>1</sup>. In addition to the goal of developing this NLG system for genetic counselors, another, broader, objective of our research is to analyze problems in communication of biomedical information to lay persons and to investigate potential solutions for use in NLG systems. For example, challenges in clinical genetics patient communication that are shared with other areas of medicine include explanation of risk and explanation of the diagnostic process and its use of evidential reasoning and statistical data.

For the biomedical content level of analysis of this corpus, we have devised a coding scheme for representing causal and probabilistic relationships among concepts at a level of abstraction that captures commonalities among genetic processes and their relationship to health. The result of applying the coding scheme to a patient letter is to model its content in a Bayesian network (BN) formalism [3,4]. Although systems using BNs have been developed for automated genetic risk analysis [5] and other biomedical decision support applications, e.g., [6–8], we know of no previous work in which a BN-based approach has been used for content annotation of text or dialogue corpora. While the initial motivation for developing the coding scheme was to support our NLG research efforts, it provides a way of encoding biomedical content in this genre that should be useful for future research on patient communication in clinical genetics. Furthermore, our general approach to representing biomedical content in a BN formalism may be useful for applications requiring (manual or automated) analysis of patient-oriented documents in other areas of medicine. In this paper, we present

the biomedical content coding scheme and a formal evaluation of the intercoder reliability of its initial tag set, which we found to be very good. Then, we briefly illustrate how we have used the coding scheme for analysis of the corpus. In addition, we propose how to automate the content annotation process and explore other uses of the coding scheme.

## 2. A Bayesian network coding scheme

### 2.1. Overview

Essentially, the coding scheme consists of a relatively small set of tags relevant to clinical genetics and constraints on their inter-relationships. Although there are more than 5000 known human genetic disorders [9], with many different direct and indirect effects, our coding scheme provides a more abstract representation that is based on commonalities among genetic processes and their relationship to health. For example, two completely different genotypes discussed in the letters, such as the genotype related to Velocardiofacial (VCF) syndrome and the genotype related to Neurofibromatosis (NF), would both be tagged as referring to the more abstract concept of *genotype*.<sup>2</sup> Also, different health problems, such as a birth defect or a developmental delay, would each be tagged as referring to the more abstract concept of *symptom*. In addition to classifying the concepts described in a letter, the encoding tracks which family member is described by a tagged phrase. For example, patient letters often discuss a known or hypothesized genotype of the mother and father of a patient. The potential causal relationships among these concepts can be represented in a graphical format as shown in Fig. 1, where arrows depict possible direct causal relationships.

The graph shown in Fig. 1 can be viewed as a simple Bayesian network (BN), where each node (depicted as a rectangle) of the network graph represents a discrete-valued random variable. The text inside the rectangles in Fig. 1 includes several types of information. The labels *genotype* and *symptom* are tags in our coding scheme denoting variable *types*. A tag appended with a numeral, e.g., *genotype-1*, is a variable *name*. (As we discuss later, due to repeated mentions of the same concept in a text, a BN variable may be identified by more than one name.) For illustrative purposes, Fig. 1 also shows the *domain* of each variable, e.g., 0, 1, or 2 copies of *VCF mutation*<sup>3</sup>. Also, the text inside a rectangle indi-

<sup>1</sup> While corpus analysis often is useful for knowledge acquisition at the beginning of an NLG project, it does not obviate the need for system evaluation by users [2].

<sup>2</sup> In clinical genetics, *genotype* refers to the two copies, or *alleles*, of a particular gene in an individual's genome. One copy is inherited from the mother, and the other from the father. In Mendelian autosomal dominant or recessive disorders, the presence of one or two abnormal copies of a gene, respectively, may result in health problems.

<sup>3</sup> i.e., zero, one, or two abnormal copies of the gene related to VCF.

متن کامل مقاله

دریافت فوری ←

**ISI**Articles

مرجع مقالات تخصصی ایران

- ✓ امکان دانلود نسخه تمام متن مقالات انگلیسی
- ✓ امکان دانلود نسخه ترجمه شده مقالات
- ✓ پذیرش سفارش ترجمه تخصصی
- ✓ امکان جستجو در آرشیو جامعی از صدها موضوع و هزاران مقاله
- ✓ امکان دانلود رایگان ۲ صفحه اول هر مقاله
- ✓ امکان پرداخت اینترنتی با کلیه کارت های عضو شتاب
- ✓ دانلود فوری مقاله پس از پرداخت آنلاین
- ✓ پشتیبانی کامل خرید با بهره مندی از سیستم هوشمند رهگیری سفارشات